# A Causal Lens for Peeking into Black Box Predictive Models: Predictive Model Interpretation via Causal Attribution

Aria Khademi,[1,2] Vasant Honavar[1,2,3,4]

[1] Artificial Intelligence Research Laboratory
[2] College of Information Sciences and Technology
[3] Department of Computer Science and Engineering
[4] Institute of Computational and Data Sciences
The Pennsylvania State University
{khademi,vhonavar}@psu.edu

August 4, 2020

## Abstract

With the increasing adoption of predictive models trained using machine learning across a wide range of high-stakes applications, e.g., health care, security, criminal justice, finance, and education, there is a growing need for effective techniques for explaining such models and their predictions. We aim to address this problem in settings where the predictive model is a *black box*; That is, we can only observe the response of the model to various inputs, but have no knowledge about the internal structure of the predictive model, its parameters, the objective function, and the algorithm used to optimize the model. We reduce the problem of interpreting a black box predictive model to that of estimating the *causal effects* of each of the model inputs on the model output, from observations of the model inputs and the corresponding outputs. We estimate the causal effects of model inputs on model output using variants of the Rubin Neyman *potential outcomes* framework for estimating causal effects from observational data. We show how the resulting *causal attribution* of responsibility for model output to the different model inputs can be used to interpret the predictive model and to explain its predictions. We present results of experiments that demonstrate the effectiveness of our approach to the interpretation of black box predictive models via causal attribution in the case of deep neural network models trained on one synthetic data set (where the input variables that impact the output variable are known by design) and two real-world data sets: Handwritten digit classification, and Parkinson's disease severity prediction. Because our approach does not require knowledge about the predictive model algorithm and is free of assumptions regarding the black box predictive model except that its input-output responses be observable, it can be applied, in principle, to *any* black box predictive model.

1

# 1 Introduction

Our ability to acquire and annotate increasingly large amounts of data together with rapid advances in machine learning have made predictive models that are trained using machine learning ubiquitous in virtually all areas of human endeavor. Recent years have seen rapid adoption of machine learning to automate decision making in many high-stakes applications such as health care [23,54,61,67,72,100,102], finance [3,18,41,88], criminal justice [70], security [38,93], education [51,75,101], and scientific discovery [13, 14, 22, 28, 63, 99]. In such high-stakes applications, the predictive models, produced by the state-of-the-art machine learning algorithms, e.g., deep learning [52], kernel methods [37], among others, are often complex black boxes that are hard to explain to users [12, 32].

**Example.** Consider a medical decision making scenario where a predictive model, e.g., a deep neural network, trained on a large database of labeled data, is to assist physicians in diagnosing patients. Given the high-stakes nature of the application, it is important that the clinical decision support system be able to explain the output of the deep neural network to the physician, who may not have a deep understanding of machine learning. For example, the physician might want to understand the subset of patient characteristics that contribute to the diagnosis; or the reason as to why diagnoses were different for two different patients, etc.

In high-stakes applications of machine learning, the ability to explain a machine learned predictive model is a prerequisite for establishing *trust* in the model's predictions. Hence, there is a growing interest in machine learning algorithms that produce *interpretable* (as opposed to black box) models as well as techniques for interpreting black box models [21, 32, 56].

## 1.1 Model Explanation and Interpretation

The nature and desiderata of explanations have been topics of extensive study in philosophy of science, cognitive science, and social sciences [43, 47, 60, 79, 80]. Arguably, predictive model *interpretation* is a necessary condition for model *explanation*. A key tool for model interpretation is *attribution* of responsibility for the model output to the model's inputs (i.e., features) [76, 90]. Hence, a large body of work has focused on methods for interpretation of black box predictive models (reviewed in [1,31,64,65]), also known as *posthoc* interpretations [31, 56]. They include methods for visualizing the effect of the model inputs on its outputs [87, 103, 104], methods for extracting purportedly human interpretable rules from black box models [5, 27, 53, 83, 91, 92], feature scoring methods that assess the importance of individual features on the prediction [4, 17, 19, 26, 29, 55, 58, 89], gradient based methods that assess how changes in inputs impact the model predictions [7, 8, 17, 85, 86], and techniques for approximating local decision surfaces in the neighborhood of the input sample via localized regression [11, 73, 82]. A shared feature of all of these model interpretation methods is that they primarily focus on how a model's inputs correlate with its outputs. As has been pointed out, they often fail to generate reliable attributions [46, 90], let alone interpretations that support explanations [1, 31, 64, 65].

## 1.2 Causal Underpinnings of Model Explanation and Interpretation

Satisfactory explanations have to provide answers to questions such as "What features of the input are responsible for the predictions?"; "Why are the model's outputs different for two individuals?" (e.g., "Why did John's loan application get approved when Sarah's was not?"). As Salmon noted, "We come to understand a phenomenon when we can explain *why* it occurred [80] (pp. 83)", where "why", has a *causal* meaning. In the words of Halpern and Pearl [33], "the role of explanation is to provide the information needed to establish *causation*" (emphasis ours) [33]. Hence, satisfactory explanations are fundamentally *causal* in nature [43, 47, 60, 62, 65, 79, 80]. Therefore, interpretations or attributions that assign responsibility for the model's output to the model's inputs, have to necessarily be causal in nature. Purely correlation based methods fail to provide causal attributions, especially in the presence of *confounders* (that is, inputs to the predictive model that causally influence both some of the other model inputs as well as model outputs). Establishing causal effect of one variable on another has to effectively cope with confounders [68]. Causal attribution is no exception.

## 1.3 Predictive Model Interpretation via Causal Attribution

Recently, [15] offered the first causal approach for interpretation of the output of a deep neural network in terms of its inputs. They offer a method for *causal attribution*, namely, assigning responsibility for the deep neural network output among its inputs. This is accomplished by estimating the causal effect of each of the model inputs on the model output. They first translate a learned deep neural network model into a *functionally equivalent* Structural Causal Model (SCM) or a Causal Bayesian Network [68] and then use the resulting Causal Bayesian Network to estimate the relevant causal effects, using variants of standard methods for causal inference using Causal Bayesian Networks.

## 1.4 Overview and Key Contributions

The only existing method [15] for causal attribution of black box predictive models suffers from at least two key limitations: (i) Since the method relies on a specific translation of a deep neural network into a Causal Bayesian Network, the method is limited in its applicability to black box predictive models that are not deep neural networks; (ii) The method requires the attribution algorithm to have access to the *structure* as well as parameters of the deep neural network. In many application scenarios, the users of the predictive model or the causal attribution algorithm lack knowledge of the internal structure of the model and its parameters, the objective function, and the algorithm used to optimize the predictive model. In such cases, the users can only observe the outputs of the model for user-supplied inputs. This further limits the applicability of the causal attribution method in [15] to settings where the deep neural network model, although complex, is *not* a black box. Against this background, we consider the causal attribution of black box predictive models.

We note that for a given deep neural network, or for that matter, any predictive model that is trained on a specific training set, a Causal Bayesian Network (CBN) that is functionally equivalent to the trained predictive model simply *cannot* include any information that is not already encoded by the trained predictive model. Hence, we conjecture that it should be possible to recover all of the information

encoded by such a Causal Bayesian Network, by observing the outputs of the corresponding predictive model on a sufficiently large sample of inputs. Thus, any causal attributions that can be estimated from the Causal Bayesian Network that is functionally equivalent to a given predictive model, can be equally well estimated from observing the outputs of the predictive model on a sufficiently large sample of inputs. If this conjecture turns out to be true, then it must be possible to leverage the state-of-the-art methods for estimating causal effects from observational data to produce causal explanations of black box predictive models and their predictions.

We reduce the model interpretation question, "Why did the predictive model generate the output $Y$ for input $X$?", to the following equivalent question: "How are the features of the model input $X$ causally related to the model output $Y$?" In other words, we reduce the task of interpreting a black box predictive model to the task of estimating, from observations of the inputs and the corresponding outputs of the model, the causal effect of each input variable or feature on the output variable. We estimate the relevant causal effects under rather mild (and quite standard) assumptions of the *Potential Outcomes* framework [34, 78] for estimating causal effects from observational data. Because unlike the only existing causal attribution method [15], we do not require the causal attribution method to have access to structure or parameters of the black box predictive model, the resulting causal attribution method can be applied, in principle, to *any* black box predictive model, so long as it can probe the model and observe the model's response to any supplied input data sample.

We demonstrate the effectiveness of the proposed approach to interpretation (via causal attribution) of black box predictive models, specifically, deep neural networks (DNN), using state-of-the-art methods for estimating causal effects from observational data, where the input variables are continuous.

The key contributions of this paper are as follows:

1. We offer the first model agnostic approach to interpretation of black box predictive models via causal attribution, that is, estimation of the causal effect of each of the model's inputs on the model's output.

2. We reduce the problem of interpretation of black box predictive models to the well-known problem of estimating the causal effects (of the model's inputs on the model's outputs) from observational data.

3. In contrast to the only existing approach to interpretation via causal attribution [15], our solution does not require the interpretation algorithm to have access to the internal structure and parameters of the black box predictive model, and hence can be applied, in principle, to *any* black box predictive model, so long as it can probe the model and observe the model's response to any supplied input data sample.

4. We show how to use the resulting causal attributions to explain the observed differences in the model's outputs in different cases, e.g., "Why did the model recommend that John's loan application be approved when Sarah's was not?"

5. We demonstrate the effectiveness of our approach to the interpretation of black box predictive models via causal attribution using DNN models trained on one synthetic data set (where the

input variables that impact the output variable are known by design) and two real-world data sets: Handwritten digit classification, and Parkinson's disease severity estimation.

The rest of the paper is organized as follows. Section 2 introduces the key definitions and the basic machinery of causal inference from observational data which we will utilize in the rest of the paper. Section 3 introduces our approach to interpretation of black box predictive models via causal attribution. Section 4 presents results of our experiments for interpreting DNN models trained on one synthetic and two real-world data sets, using several state-of-the-art methods for causal effects estimation from observational data. Section 5 concludes with a brief summary, a discussion of the related work, some caveats regarding the applicability of the proposed approach to interpretation of black box predictive models, and some promising directions for further research.

# 2 Preliminaries

## 2.1 Causal Effects

The central problem in causal inference is determining whether, and how, a change in a treatment $T$ (e.g., surgery) leads to a change in some outcome $Y$ (e.g., health status).

We will introduce the key notions when both the treatment and outcome are binary (for simplicity) before proceeding to consider the setting where the treatments are continuous-valued.

## 2.2 Estimating Causal Effects of Discrete Treatments

Let $Y_i^{(t)}$ and $Y_i^{(t')}$ be the *potential outcomes* when an individual $i$, is exposed to treatments $T = t$ and $T = t'$, respectively. The causal effect of $T$ on $Y$ is gauged with contrasting $Y_i^{(t)}$ and $Y_i^{(t')}$. An estimand of interest is the average causal effect, which is defined as follows if the treatment is binary:

**Definition 1. (Average Causal Effect (ACE))** *Consider a population of individuals each with the potential outcomes $Y_i^{(t)}$ and $Y_i^{(t')}$. The* Average Causal Effect *of $T$ on $Y$ is defined as:*

$$ACE_T^Y = \mathbb{E}[Y_i^{(t)} - Y_i^{(t')}] = \mathbb{E}[Y_i^{(t)}] - \mathbb{E}[Y_i^{(t')}]. \tag{1}$$

Because for each individual $i$, the random variable $T$ must either take the value $T = t$ or $T = t'$ but not both, only one of the potential outcomes $Y_i^{(t)}$ or $Y_i^{(t')}$ is observable. For example, we observe the effect of a surgery on the individual's health status (e.g., cured), but cannot observe the effect of not having done surgery for the *same individual*. The observable outcome is called the factual outcome and the unobservable outcome is called the counterfactual outcome. Counterfactual inference requires us to estimate the outcome that *would have been observed*, had the treatment variable been assigned a value that is different from its observed value. Within the potential outcomes framework, the counterfactual outcomes are estimated from the observed outcome(s) for individual(s) of the opposite group that are *most similar* to the individual under consideration [34]. This procedure may become unreliable in high-dimensional spaces [2]. To cope with this problem, we can use modern representation

5

learning techniques [10, 95] to map the data to a low-dimensional latent space before estimating the counterfactual outcomes [42, 57, 81], or employ targeted maximum likelihood methods (TMLE) [97] for counterfactual inference from observational data.

## 2.3 Causal Effects Under Continuous Treatment

If the treatment is continuous, the potential outcomes can still be written as $Y_i^{(t)}$ where $t$ can take any value in a continuum, e.g., $t \in [t_{min}, t_{max}]$ and one can contrast two values of the potential outcomes for any $t$ and $t'$. The unit-level counterfactual outcomes are unobservable and an estimand of interest is $\mathbb{E}[Y_i^{(t_i)}]$, the average treatment effect. The potential outcomes framework [78] offers a variety of estimators to estimate the average treatment effect in a continuous treatment regime, e.g., covariate balancing propensity score and its non-parametric counterpart [25], propensity score weighting using generalized boosted models [107], optimization based weighting [108], inverse probability of treatment weighting in generalized linear models [74], propensity score weighting using the super learner [71, 96] (see [30] for a review and Section 3.1 for details).

## 2.4 Assumptions

The potential outcomes approach to counterfactual inference [34] makes a set of assumptions:

1. Consistency: The potential outcome of any individual $i$ that has been exposed to a treatment $t$ is the realized outcome for that individual if the individual was treated with $t$. In other words, $T_i = t \implies Y_i^{(t)} = Y_i$;

2. Positivity: In the case of discrete treatments, each possible treatment has non-zero probability; and in the case of continuous treatments, the conditional density of treatment given the covariates is non-negative for all covariates;

3. Stable Unit Treatment Value Assumption (SUTVA): The potential outcomes of any individual do not depend on the treatment assigned to other individuals; All treated (untreated or controlled) individuals receive the same version of treatment (control);

4. Unconfoundedness: There are no unobserved confounders (i.e., variables that causally impact both the treatment and the outcome) and hence, adjusting for the observed confounders eliminates any bias in comparing the treated and untreated individuals – an assumption that is untestable from observational data alone.

Unconfoundedness and positivity (together referred to as strong ignorability) imply that the causal effect of the treatment is identifiable from observational data [34]. In this paper, we will assume that the assumptions generally hold. See Section 5 for discussion of how some of these assumptions can be relaxed if needed.

# 3   Predictive Model Interpretation via Causal Attribution

Let $D = \{\mathbf{X}_i, Y_i\}_{i=1}^n$ be a set of data samples from $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space, or the domain of feature vector $\mathbf{X}$, and $\mathcal{Y}$ is the domain of the outcome $Y$. Let $f_\mathbf{W} : \mathcal{X} \to \mathcal{Y}$, be a predictive model trained on $D$, so as to minimize some objective function, e.g., a suitable measure of error of the model's predictions $h(\mathbf{X}_i)$, for $\mathbf{X}_i$ relative to the desired output $Y_i$, over the training data, using a suitable learning algorithm $L$. The resulting predictive model $f_\mathbf{W}$, given an input $\mathbf{X} \in \mathcal{X}$, produces an output $Y \in \mathcal{Y}$.

Recall that our goal is to interpret or explain a black box predictive model $f_\mathbf{W}$. We aim to do so by attributing the responsibility for the model output to the model's inputs (i.e., features) [76, 90]. We focus on interpretation via causal attribution in the setting where the interpretation algorithm has no knowledge about the internal structure, or the parameters of the black box predictive model $f_\mathbf{W}$. Thus, if $f_\mathbf{W}$ is a deep neural network, the causal attribution algorithm can be expected to have no knowledge about the architecture of the deep neural network, the objective function used to optimize its parameters, or the parameters of the trained network.

We reduce the problem of interpreting the predictive model $h$ to the problem of determining the *causal effect* of each of the features of $\mathbf{X}$ on $\mathbf{Y}$ from observations of a sufficiently large data set of sample model inputs and the corresponding model outputs. We aim to answer the following question: Why did $f_\mathbf{W}$ generate $Y = f_\mathbf{W}(\mathbf{X})$ as the output for input $\mathbf{X}$? In other words, what is the causal effect of each of the features of $\mathbf{X}$ on $Y$? To answer this question, for each feature $X_j$ of $\mathbf{X}$, we designate $X_j$ as the "treatment" and estimate its average causal effect on $Y$. Consider the set of data samples $\{\mathbf{V_i}, T_i, Y_i\}_{i=1}^n$ where $\mathbf{V_i} = \mathbf{X_i} \setminus T_i$. For each choice of $T$, we estimate the average causal effect of $T$ on $Y$, i.e., $ACE_T^Y$, yielding a vector of causal effects $\mu = \{\mu_1, \ldots, \mu_m\}$ where each $\mu_j, j = 1, \ldots, m$ denotes the estimated causal effect of feature $X_j, j = 1, \ldots, m$ on $Y$. We will use state-of-the-art methods for estimating the causal effect of each input on the output of the predictive model from the observed input-output samples.

**Definition 2. (Causal Attribution)** *Given a predictive model $h : \mathcal{X} \to \mathcal{Y}$ trained on a data set $D$, causal attribution of the output $Y$ is $\mathbf{A} = \{a_1, \ldots, a_m\}$, where each $a_j = \mu_j$, is the causal effect of the feature $X_j$ on $Y$ iff $\mu_j$ is statistically significantly non-zero, and $a_j = 0$ otherwise.*

The non-zero causal effects included in the causal attribution are identified by testing the null hypothesis $H_0 : ACE_{x_j}^y = 0$, $\forall j \in \{1, \ldots, m\}$, where $x_j$ denotes the model inputs and $y$ the model output. If $H_0$ is rejected at a chosen level of statistical significance, then $x_j$ is causally responsible for output $y$ and the degree of its influence is gauged by the estimated causal effect.

## 3.1   Estimating Causal Effects

We use potential outcomes based state-of-the-art methods for estimating causal effects of continuous treatments on outcomes [30]. The mechanism of the methods in this paper (described in detail below) involves two steps:

1. Estimation of weights, $w_i = f(T_i, \mathbf{X_i}, \epsilon_i)$, for each data sample to ideally achieve independence between their features and the treatment and thus, deconfounding (or balancing) the treatment and outcome;

2. Modeling the outcome $Y_i = g(T_i, \mathbf{X_i}, \epsilon_i)$, as a function of the *weighted* data samples.

If either the treatment model $f(\cdot)$ or the outcome model $g(\cdot)$ are correctly specified, the estimated result is an unbiased estimate of the average causal effect of the treatment on the outcome. The methods that we use have been shown to be effective in a variety of causal effect estimation applications [9, 16, 25, 45, 48, 106] and are listed as follows:

1. **Covariate Balancing Propensity Score (CBPS)** [25] which is an estimate of the generalized propensity score (GPS) or the conditional propensity density, defined as $r(t, x) = f(T_i = t_i \,|\, \mathbf{X}_i = \mathbf{x}_i)$, while achieving covariate balance. Generalized propensity score is a generalization, to continuous treatment regimes [35, 40], of the propensity score [77], a well-established and very commonly used distance measure for counterfactual estimation in binary treatment regimes, which is defined as $Pr(T_i = 1 \,|\, \mathbf{X}_i = \mathbf{x}_i)$. CBPS estimates the GPS under covariate balancing constraints to maximize covariate balance, thereby avoiding the need for an iterative model fitting process until the desired balance is achieved.

2. **Non-Parametric Covariate Balancing Propensity Score (NPCBPS)** [25] which is a non-parametric approach that maximizes the likelihood of the observed data under covariate balancing constraints for computing the GPS weights.

3. **Propensity Score Weighting Using Generalized Boosted Models (PSWGBM)** [107] which estimates the propensity scores using generalized boosted models. In this method, we use the mean of the Spearman correlation statistic to maximize balance between data points of different treatments.

4. **Optimization-Based Weighting (OPTWEIGHT)** [108] which solves a quadratic optimization problem constrained to achieve covariate balance (we used a million iterations and set the absolute difference in means of the weighted features $\delta \leq 0.1$ as the convergence criterion).

5. **Inverse Probability of Treatment Weighting in Generalized Linear Models (IPTW)** [74] which estimates the propensity scores using generalized linear models and weights each data sample according to the inverse of the estimated propensity score.

6. **Propensity Score Weighting Using Super Learner (SUPER)** [71, 96] which is designed to be robust to misspecification of the treatment model, the outcome model, or both. It estimates the propensity score based on an optimized weighted combination of an ensemble of candidate prediction models optimized using cross validation.

# 4  Experiments and Results

We proceed to report results of our experiments with the proposed methods for interpretation via causal attribution of black box predictive models. To facilitate direct comparisons with [15], we used the trained neural network as the model to be interpreted, and the inputs and outputs of neural network as observational data for estimating the causal effect of each input of the neural network on its output. We also report results on an additional real-world data set and the trained DNN from [4].

## 4.1  Data Sets and DNN Models Used

### 4.1.1  Synthetic data set

We used a synthetic data set from [15], generated following [36] where a data sample consists of 12 features, $f_1, \ldots, f_{12}$, sampled from the normal distribution according to the following procedure: $\forall i \in \{4, \ldots, 12\} : f_i \sim \mathcal{N}(0, 0.2)$, and

1. either (with probability 0.5) $\forall j \in \{1, 2, 3\}$: $f_j \sim \mathcal{N}(1, 0.2)$ and set the class label to 1,

2. or (with probability 0.5) $\forall j \in \{1, 2, 3\}$: $f_j \sim \mathcal{N}(-1, 0.2)$ and set the class label to 0.

The data generation procedure ensures that only the first 3 features are responsible for class labels. We used the 3-layer neural network used in the experiments of [15]. We generated a 1000 data samples from the synthetic data generator described above and obtained the predicted outputs of the neural network on each of the data samples. We used the resulting data for causal attribution.

### 4.1.2  MNIST data set

We used a convolutional neural network (CNN) trained on the MNIST training data to classify the handwritten digits of $0, \ldots, 9$. We ran each test image with 784 features ($28 \times 28$ pixels) through the trained CNN and recorded the predicted class labels. For ease of interpretation, we trained an Auto-Encoder (AE) on the test data to reduce the dimensions from 784 to 10 latent variables, $Z_0, \ldots, Z_9$, and labeled each 10-dimensional feature vector with the corresponding CNN-predicted class label for the $28 \times 28$ or 784 pixel image. We used the resulting data for causal attribution of the CNN output to the 10 latent variables. [1]

### 4.1.3  Parkinson's Disease (PD) Telemonitoring data set

The PD telemonitoring data set [94] provides age, gender, and 16 (at-home) biomedical voice measurements (processed from voice recordings) obtained from PD patients (see [94] for details) along with their scores on the Universal Parkinson's Disease Rating Scale (UPDRS) which measures the severity

---

[1]Note that we could have performed the same analysis with 784 pixel variables, but we chose the 10 latent variables setup used here for ease of visualization, and to demonstrate the power of the proposed method to perform causal attribution with respect to *any* variables that are derived from the input variables (which in this case are the 10 latent variables).

and progression of the Parkinson's disease. We used the DNN trained by [4] on the PD data and used it to predict UPDRS for each PD data sample. We used the inputs and outputs of the DNN for each of the PD data samples for causal attribution.

## 4.2 Causal Attribution of Black Box Models

In keeping with our goal of interpreting black box models, we used *only* the observed inputs and outputs of the model (and not any information related to the structure, parameters, or the algorithm used to optimize the model parameters) for causal attribution of the model output relative to the model inputs. To estimate the causal effects of the model inputs on the model output, we used the R package WeightIt [30] (version 0.7.1). Unless otherwise noted, we used the default parameter settings for each method. In identifying non-zero causal effects, we set the significance level $\alpha = 0.05$.

## 4.3 From Causal Attributions to Contrastive Explanations

We also explored how the causal attributions of model output with respect to model inputs can be used to explain why the model output for a specific input data pattern differs from that for another. Specifically, we focus on the features that have large causal attributions for the model prediction, and examine how the two data samples being contrasted differ with respect to the values of those features.

## 4.4 Experimental Results

### 4.4.1 Synthetic data set

We estimated the causal effects of all features of the data set on the class label predicted by the neural network and report the results in Table 1.

Upon testing the null hypothesis (that the causal effect of each feature is 0), we find, from all of the causal effect estimation methods, that the output of the neural network is causally impacted by only the features $f_1, f_2, f_3$ and not $f_4, \ldots, f_{12}$. This finding is in agreement with the design of the simulated data generator (which ensures that only the features $f_1, f_2, f_3$ determine the class label) and the results reported in [15, 90].

**Example:** Consider, for example, the neural network constructed from the simulated data where we found the features $\{f_1, f_2, f_3\}$ to have large causal effects on the model's predictions. Now suppose a user wants to understand why the model outputs class label $0$ for sample

- $\mathbf{S}_1 = $ [-0.89, -1.11, -0.85, -0.33, -0.47, 0.23, -0.20, 0.13, -0.17, 0.35, -0.22, 0.04],

and class label 1 for sample

- $\mathbf{S}_2 = $ [0.81, 0.95, 1.11, 0.01, 0.12, -0.22, 0.23, 0.18, 0.10, 0.18, -0.14, -0.02].

The answer to the user's question can be obtained by (i) noting that only the first three features, $\{f_1, f_2, f_3\}$, have non-zero causal attributions and (2) recognizing that there is a clear shift in the values of the features $\{f_1, f_2, f_3\}$ from the mean of the features for class $0$ samples towards the mean for class $1$ samples.

10

Table 1: Estimates and p-value significance of causal effects of input features on the outputs of the neural network trained using the synthetic data set.

| Feature | CBPS | | NPCBPS | | PSWGBM | | IPTW | | OPTWEIGHT | | SUPER | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Est. | P | Est. | P | Est. | P | Est. | P | Est. | P | Est. | P |
| f1 | 1.69 | <0.01 | 1.36 | <0.01 | 0.62 | <0.01 | 1.81 | <0.01 | 2.53 | <0.01 | 1.53 | <0.01 |
| f2 | 1.58 | <0.01 | 1.20 | <0.01 | 1.51 | <0.01 | 1.60 | <0.01 | 2.67 | <0.01 | 1.24 | <0.01 |
| f3 | 1.83 | <0.01 | 1.00 | <0.01 | 0.54 | <0.01 | 1.93 | <0.01 | 2.26 | <0.01 | 1.76 | <0.01 |
| f4 | 0.00 | 0.96 | 0.00 | 0.96 | 0.04 | 0.63 | -0.01 | 0.94 | 0.00 | 0.97 | 0.02 | 0.80 |
| f5 | 0.07 | 0.42 | 0.07 | 0.43 | 0.09 | 0.33 | 0.07 | 0.43 | 0.07 | 0.41 | 0.07 | 0.44 |
| f6 | -0.11 | 0.21 | -0.11 | 0.20 | -0.04 | 0.67 | -0.11 | 0.19 | -0.11 | 0.21 | -0.07 | 0.39 |
| f7 | 0.16 | 0.06 | 0.16 | 0.05 | 0.17 | 0.05 | 0.16 | 0.05 | 0.16 | 0.06 | 0.14 | 0.09 |
| f8 | -0.04 | 0.67 | -0.04 | 0.66 | -0.03 | 0.76 | -0.03 | 0.71 | -0.03 | 0.69 | -0.05 | 0.54 |
| f9 | -0.04 | 0.67 | -0.04 | 0.67 | 0.03 | 0.69 | -0.04 | 0.68 | -0.03 | 0.70 | -0.02 | 0.83 |
| f10 | 0.15 | 0.09 | 0.15 | 0.08 | 0.16 | 0.06 | 0.15 | 0.08 | 0.15 | 0.08 | 0.15 | 0.08 |
| f11 | 0.07 | 0.42 | 0.08 | 0.39 | 0.02 | 0.81 | 0.07 | 0.44 | 0.07 | 0.41 | 0.06 | 0.52 |
| f12 | -0.19 | 0.04 | -0.19 | 0.04 | -0.14 | 0.11 | -0.19 | 0.04 | -0.19 | 0.04 | -0.12 | 0.18 |

### 4.4.2 MNIST data set

Using all causal effect estimation methods described in Section 3.1, we estimated the causal effect of latent variables (of test data) $Z_0, \ldots, Z_9$, obtained using the AE, on the class label $C_k : \forall k \in \{0, \ldots, 9\}$, predicted by the CNN. We obtained causal attributions of the predicted class wih respect to the latent variables. To validate our results, we proceeded with two steps: (1) We intervened on each of the causal latent variables by manually setting their value to zero and reconstructed the image. If a latent variable has a large non-zero causal attribution for the predicted label, we expect the reconstructed image to deviate from its original appearance; and (2) We intervened on all of the non-causal latent variables, i.e., those with low causal attributions, by setting their value to zero and reconstructed the image using only the remaining variables. We expect the resulting reconstruction to be similar to the original image.

We show the results of our experiments in Figure 1 (best viewed in color). We observe that intervening on and zeroing out each of the identified latent variables with large causal attribution for the predicted CNN label, indeed results in a significant distortion of the reconstructed image relative to the original image. The distorted images can be viewed in columns (b) and (c), while the original image can be viewed in column (a) of Figure 1. On the other hand, images reconstructed using only the identified causal latent variables end up being similar to the original image. The images reconstructed using only the identified causal latent variables are shown in column (d) of Figure 1. Interestingly, we observed that all of the causal effect estimation methods consistently agreed with each other in identifying the causal latent variables in this experiment. These results demonstrate that our method is indeed effective in correctly interpreting black box predictive models via causal attribution.

**Example:** Suppose a user wants to know why a specific image is classified as the digit 3. The answer to
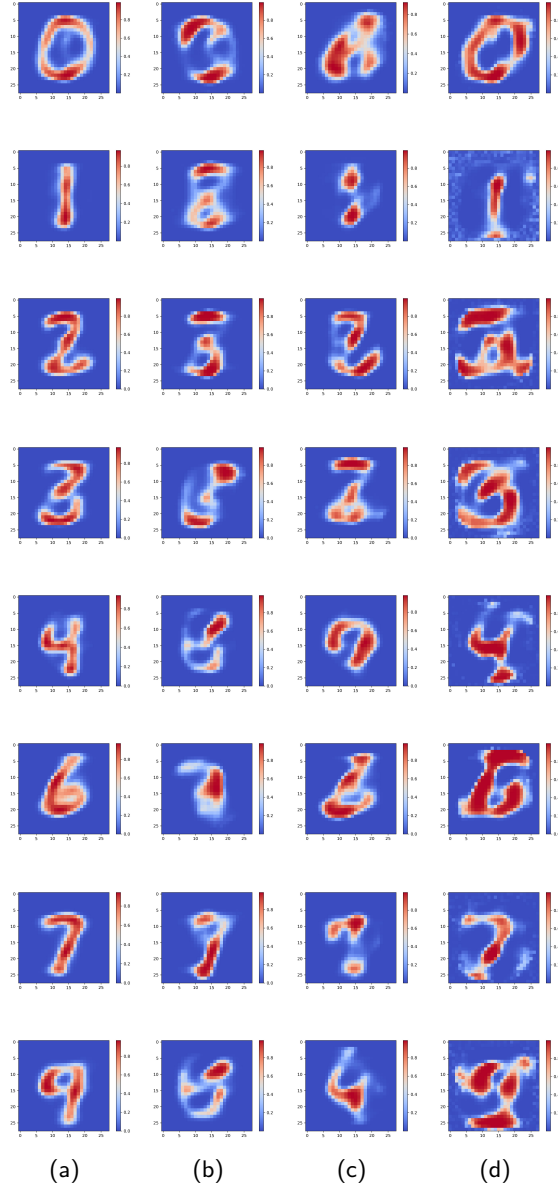
Figure 1: Results of our experiments on the MNIST data set. (a) Images reconstructed using the Auto-Encoder. (b) and (c) Images reconstructed after intervening on each causal latent variable. (d) Images reconstructed using only the causal subset of latent variables, having set all other latent variables to zero.

Table 2: Estimates and p-value significance of the causal effect of each feature on the predicted UPDRS by the deep neural network on the PD Telemonitoring data set.

| | CBPS | | NPCBPS | | PSWGBM | | OPTWEIGHT | | IPTW | | SUPER | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature | Est. | P | Est. | P | Est. | P | Est. | P | Est. | P | Est. | P |
| Age | 16.12 | <0.01 | 15.20 | <0.01 | 15.65 | <0.01 | 16.81 | <0.01 | 15.56 | <0.01 | 15.54 | <0.01 |
| Gender | -1.48 | <0.01 | -1.76 | <0.01 | -1.51 | <0.01 | -2.37 | <0.01 | -1.83 | <0.01 | -1.83 | <0.01 |
| Jitter (%) | 62.75 | 0.12 | 16.99 | 0.70 | 12.20 | 0.77 | -11.30 | 0.78 | -5.07 | 0.68 | -6.80 | 0.58 |
| Jitter (Abs) | -13.21 | <0.01 | -38.50 | <0.01 | -16.84 | 0.02 | -60.94 | <0.01 | -36.24 | <0.01 | -36.16 | <0.01 |
| Jitter:RAP | -2713.96 | <0.01 | 23826.27 | <0.01 | -3561.39 | 0.21 | 3054.57 | 0.38 | -4775.70 | <0.01 | -4803.89 | <0.01 |
| Jitter:PPQ5 | 29.17 | 0.02 | 21.58 | 0.39 | -31.16 | 0.19 | 57.19 | 0.04 | -43.32 | <0.01 | -43.54 | <0.01 |
| Jitter:DDP | 2536.11 | 0.01 | 2307.36 | 0.49 | 3591.26 | 0.20 | -3029.24 | 0.38 | 4850.61 | <0.01 | 4879.72 | <0.01 |
| Shimmer | 117.01 | <0.01 | -13.76 | 0.74 | 49.15 | 0.10 | -102.39 | 0.03 | 62.88 | <0.01 | 59.94 | <0.01 |
| Shimmer:dB | 0.61 | 0.95 | -16.65 | 0.40 | -17.11 | 0.35 | -8.79 | 0.65 | 36.93 | <0.01 | 37.72 | <0.01 |
| Shimmer:APQ3 | 5119.44 | <0.01 | -1094.98 | 0.53 | 1628.27 | 0.45 | 455.96 | 0.72 | -940.25 | 0.07 | -938.57 | 0.07 |
| Shimmer:APQ5 | -34.58 | <0.01 | -43.04 | 0.07 | -16.91 | 0.30 | -75.33 | <0.01 | -44.94 | <0.01 | -45.67 | <0.01 |
| Shimmer:APQ11 | 32.02 | 0.05 | 48.73 | 0.03 | 12.18 | 0.38 | 62.88 | <0.01 | 22.24 | <0.01 | 22.67 | <0.01 |
| Shimmer:DDA | -3723.52 | <0.01 | 210.39 | 0.91 | -1674.89 | 0.44 | -348.83 | 0.79 | 1065.98 | 0.04 | 1064.20 | 0.04 |
| NHR | 0.07 | 0.99 | 26.65 | 0.01 | -5.42 | 0.55 | -36.64 | <0.01 | -8.15 | 0.09 | -8.00 | 0.10 |
| HNR | -23.40 | <0.01 | -29.40 | <0.01 | -18.58 | <0.01 | -28.17 | <0.01 | -35.13 | <0.01 | -35.43 | <0.01 |
| RPDE | 3.30 | 0.19 | 3.77 | 0.14 | -1.11 | 0.65 | 3.13 | 0.24 | -0.99 | 0.69 | -1.05 | 0.67 |
| DFA | -7.14 | <0.01 | -6.59 | <0.01 | -9.85 | <0.01 | -2.97 | 0.04 | -7.09 | <0.01 | -7.10 | <0.01 |
| PPE | 12.72 | <0.01 | 18.00 | <0.01 | 21.17 | <0.01 | 17.02 | <0.01 | 10.15 | <0.01 | 9.96 | <0.01 |

this question follows from the fourth row of Figure 1, which shows that the variables $Z_2$ and $Z_7$ (which we identified to be the responsible causal latent variables for class label 3) have large causal attributions for that prediction outcome. If we set the value of any of these variables to zero (shown in columns (b) and (c) of the figure respectively for $Z_2$ and $Z_7$ for the case of digit 3), the image will no longer look like the digit 3. If we set the values of all of the latent variables except $Z_2$ and $Z_7$ to zero, we note that even though the image is distorted, we can recognize digit 3. These results show that the specific image was classified as digit 3 because of the large causal effect of the latent variables $Z_2$ and $Z_7$.

### 4.4.3   Parkinson's Disease Telemonitoring data set

Table 2 shows our estimates of the causal effects of all 18 features of the data set on the UPDRS predicted by the neural network.[2] Interestingly, we observe that age and gender are found to have large causal effects on the predicted UPDRS score. Specifically, predicted UPDRS increases with age, and is lower in females as compared to males. Interestingly, we find that biomedical voice measures, Jitter (Abs), HNR, DFA, and PPE have significant causal effects on predicted UPDRS score (with mostly a general agreement among the different causal effect estimation methods). We note that these findings are consistent with the relevant literature [20, 94], thus supporting the notion that causal attribution of predictions of black box models can provide useful insights into data.

---

[2]We note that gender is the only discrete attribute in this data set and we have treated it as binary in our experiments. This means that the change in output is proportionate to $\exp(\hat{\mu})$ in the case of gender, where $\hat{\mu}$ is the estimated causal effect of gender on UPDRS.

# 5    Summary and Discussion

## 5.1    Summary

Predictive models trained using machine learning are extensively used across a wide range of high-stakes applications including health care, security, criminal justice, finance, and education. Such utilization imposes a pressing need for effective techniques that can explain the predictive models and their predictions. We have addressed this problem in settings where the predictive model is a *black box*, meaning that users can only observe the response of the model to various inputs, but have no information about the structure of the predictive model, its parameters, or the objective function and the algorithms that are used to optimize the model. We reduce the problem of interpretation of black box predictive models to the problem of estimating the causal effects of model inputs on model outputs from observations of outputs of the model for a sufficiently large and diverse sample of inputs. To the best of our knowledge, we offer the first model agnostic solution to interpretation of black box predictive models via causal attribution. In contrast to the only existing approach to interpretation via causal attribution [15], our solution does not require the interpretation algorithm to have access to the internal structure, parameters, or the objective function and the algorithm used to optimize the black box predictive model. Hence, the approach can be applied, in principle, to *any* black box predictive model, so long as it can probe the model and observe the model's output for any supplied input data sample. We estimate the causal effects of model inputs on model output using state-of-the art variants of the Rubin Neyman *potential outcomes* framework for estimating causal effects from observational data. The proposed solution works for both discrete as well as continuous valued inputs. We show how the resulting causal attribution of responsibility for model predictions to a subset of the inputs, can be used to explain the observed differences in the model's outputs in different cases, e.g., "Why did the model recommend that John's loan application be approved but Sarah's was not?" We demonstrate the effectiveness of our approach to the interpretation of black box predictive models via causal attribution using deep neural network models trained on one synthetic data set (where the input variables that impact the output variable are known by design) and two real-world data sets: Handwritten digit classification, and Parkinson's disease severity prediction.

## 5.2    Related Work

The nature and desiderata of explanations have been widely studied in philosophy of science, cognitive science, and social sciences [43, 47, 60, 79, 80]. The ability to interpret a predictive model is a necessary condition for being able to explain it. A dominant approach to model interpretation involves *attribution* of responsibility for the model output to the model's inputs (i.e., features) [76, 90].

Shapley values [84] offer an alternative approach to explaining predictive models [4, 19, 55, 58, 89]. In this setting, a prediction is explained by assuming that each feature of the data sample is a "player" in a game where the payout is the difference between the predicted output for that feature, and the average predicted output over the entire data set. Techniques from game theory are used to fairly distribute the "payout" among the features [84]. The Shapley value of a feature is the average marginal contribution of a feature value across all possible coalitions. While Shapley values provide an intuitive approach to

scoring features, because of the combinatorial nature of the task, the computational cost of calculating Shapley values becomes prohibitive when the number of features is large [4]. Recent work has pointed out that mathematical problems arise when Shapley values are used for feature importance and that the solutions to mitigate these necessarily induce further complexity [49]. In light of this, causal attribution methods introduced in this paper offer an attractive alternative to Shapley values for assessing feature importance.

As previously mentioned, much work has focused on attribution methods for interpreting predictive models (reviewed in [1,31,64,65]), also known as *post hoc* interpretations [31,56]. Such methods include techniques for visualizing the effect of the model inputs on its outputs [53, 87, 103, 104], methods for extracting purportedly human interpretable rules from black box models [5,27,83,91,92], feature scoring methods that assess the importance of individual features on the prediction [4, 17, 19, 26, 29, 55, 58, 89], gradient based methods that assess how changes in inputs impact the model predictions [7,8,17,85,86], and techniques for approximating local decision surfaces in the neighborhood of the input sample via localized regression [11, 73, 82]. A shared feature of all of these model interpretation methods is that they focus primarily on how a model's inputs correlate with its outputs. They often fail to generate reliable attributions [46, 90], let alone interpretations that support explanations [1, 31, 64, 65].

On the other hand, the causal underpinnings of explanations have been well-recognized in the philosophy of science, cognitive science, and social sciences literature [43, 47, 60, 79, 80, 98]. There is a growing realization that explanations in general, and those of predictive models and their predictions in particular, have to be necessarily causal [33, 60, 65]. Purely correlation based methods fail to adjust for the effect of confounders and hence are incapable of providing causal attributions. In contrast, causal inference methods [34, 68] offer powerful machinery for causal attribution.

Our work is inspired by the seminal work of [15] who offered the first causal attribution method for deep neural networks by estimating the causal effect of each of the network inputs on the network output. They achieve this by first translating the deep neural network into a functionally equivalent Causal Bayesian Network [68] which is then used to calculate the relevant causal effects. A key limitation of their method is that it requires the causal attribution algorithm to have access to the structure as well as parameters of the trained deep neural network. Our approach overcomes this limitation by observing that the Causal Bayesian Network obtained from a deep neural network, or for that matter, any predictive model trained on a given data set, fundamentally cannot provide any information that is not available in the data used to train the network. Hence, assuming that the trained network is sufficiently accurate on the training data, it should be possible to estimate the causal effects of the model inputs on its outputs using state-of-the-art methods for estimating causal effects from observational data.

We would be amiss if we did not mention possible connections between our approach and the well-known Partial Dependence Plots (PDP) [26]. PDPs offer a technique for visualizing the marginal effect of features on the output of a predictive model. Suppose that $g(x)$ is a predictive model trained using features in $\mathbf{X}$. Suppose we are interested in the effect of a subset $\mathbf{X}_S$ (where the subscript "S" refers to a subset of features in $\mathbf{X}$) on the output of $g(x)$. The partial dependence of $g(x)$ on $\mathbf{X}_S$ is defined as:

$$g_S(x_S) = \mathbb{E}_{\mathbf{X}_C}[g(x_S, \mathbf{X}_C)] = \int g(x_S, x_C) dP(x_C), \qquad (2)$$

where $\mathbf{X}_C = \mathbf{X} \setminus \mathbf{X}_S$. PDP shows the marginal effect of a feature or a set of features on the predictions. Visualizing PDP is feasible for only one or two features at a time. The recent work of [106] has shown that when a causal structure of the domain is available, PDP can be used, under some conditions, to measure the causal effect of some feature(s) on the prediction. However, such a causal interpretation of PDP requires a structural causal model to determine whether a particular subset of features satisfy the necessary conditions for causal interpretation. It would be interesting to explore the relationship between PDP and its extensions, e.g., individual conditional expectation [29], and the causal attribution of predictions to the features introduced in this paper.

It is worth noting that there is a body of recent work on algorithmic fairness [39, 44, 45, 50, 66, 105] which has shown that the problem of determining whether or not a predictive model is discriminatory with respect to a protected attribute, e.g., gender, race, can be reduced to the problem of determining whether the protected attribute has a causal effect on the output of the predictive model. A predictive model is deemed non-discriminatory with respect to a protected attribute if the causal effect of the protected attribute on the output of the predictive model is negligible (ideally 0). Clearly, methods such as those introduced in this paper for black box predictive model interpretation via causal attribution can be applied to answer algorithmic fairness questions, and help ensure that such models do not become instruments of unfair discrimination on the basis of gender, race, etc. We further note that most of the work on causal criteria for algorithmic fairness have focused on protected attributes that are binary or categorical. The causal attribution method introduced in this paper can cope with protected attributes that are ordinal or continuous e.g., age.

## 5.3 Limitations and Caveats

The proposed approach to causal attribution relies on state-of-the-art methods for estimation of causal effects from observational data, and as such, the obtained causal attributions depend on the estimated causal effects. These estimates, in turn, depend on the accuracy of the specific method used to estimate the causal effect of each input of the predictive model on the model's output. Furthermore, sometimes the different models disagree on the magnitude as well as the sign of the estimated causal effect. Under the circumstances, in practice, it makes sense to trust the causal attributions where the corresponding causal effects estimated by the different methods are largely in agreement with each other, and question the causal attributions where the corresponding causal effect estimates disagree with each other.

The proposed approach to black box predictive model interpretation via causal attribution assumes that the causal effect of each of the model inputs on the model output can in fact be estimated using only the observations of the model's input-output behavior on a sufficiently large number of data samples. Estimation of causal effects from observational data within the potential outcomes framework relies on the strong ignorability assumption which can be roughly paraphrased as saying that any causal relationships between the potential outcomes and the treatment are fully explained by the observables, i.e., that there are no unmeasured confounders [34]. Strong ignorability is the counterpart of the back-door criterion [69], a graphical criterion that if satisfied by a set of observed variables $\mathbf{Z}$ of a structural causal model, is a sufficient condition for identifiability of the causal effect of a treatment $T$ on an outcome $Y$ from observational data. It requires that no element of $\mathbf{Z}$ is a graphical descendent of $T$

and $\mathbf{Z}$ blocks all *back door* paths from $T$ to $Y$. Such a set $\mathbf{Z}$ of observed variables correspond precisely to the set of confounders that need to be controlled for in estimating the causal effect of $T$ on $Y$ from observational data. However, the backdoor criterion, or equivalently, strong ignorability, cannot be verified or refuted from *only* observational data [68], and requires a structural causal model. In general, such a structural causal model has to be either assumed, perhaps based on prior knowledge, or derived, e.g., by translating a deep neural network or other predictive model trained on an observational data set [15]. In either case, the validity of such a structural causal model cannot be verified from observational data alone.

It is possible to cope with violations of the strong ignorability assumption in some settings. For example, one can cope with *unmeasured confounders*, whenever possible, by identifying the confounders (using background knowledge) and adjusting for them [34,68]. A second approach is to use instrumental variables [6]. A third approach is to estimate bounds on the causal effects instead of the causal effects themselves [59]. Development of methods that cope with violations of the strong ignorability assumption in the presence of hidden confounders remains an active area of research [24].

## 5.4   Future Work

Causal attribution is a small, albeit important and necessary step towards explaining complex predictive models trained using machine learning. However, in addition to being causally grounded, explanations have to be selective, formulated at the appropriate level of abstraction, context-specific, concise, and framed relative to the knowledge of the explainee and the purpose of the explanation. Almost all of the work on explaining predictive models trained using machine learning fall short of these objectives. Hence, there is much room for (i) developing more advanced approaches to producing explanations from predictive models, data, background knowledge and assumptions, context, etc.; (ii) relaxing some of the limiting assumptions of the potential outcomes framework to broaden the scope of applicability of the causal attribution methods introduced in this paper; (iii) developing better criteria for evaluating and comparing alternative explanations; and (iv) approaches to integrating explanatory machinery into robust, interactive, transparent, accountable, and trustworthy human-AI systems.

# References

[1] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[2] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pages 420–434. Springer, 2001.

[3] S. Albanesi and D. F. Vamossy. Predicting consumer default: A deep learning approach. Technical report, National Bureau of Economic Research, 2019.

[4] M. Ancona, C. Oztireli, and M. Gross. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 272–281, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[5] R. Andrews, J. Diederich, and A. B. Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6):373–389, 1995.

[6] J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

[7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015.

[8] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MÃžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

[9] P. Barlow, R. Loopstra, V. Tarasuk, and A. Reeves. Liberal trade policy and food insecurity across the income distribution: an observational analysis in 132 countries, 2014–17. *The Lancet Global Health*, 8(8):e1090–e1097, 2020.

[10] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[11] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657, 2020.

[12] J. Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, 2016.

[13] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

[14] D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, and J. J. Collins. Next-generation machine learning for biological networks. *Cell*, 173(7):1581–1592, 2018.

[15] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian. Neural network attributions: A causal perspective. In *International Conference on Machine Learning*, pages 981–990, 2019.

[16] G. Chen, D. Zeng, and M. R. Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516):1509–1521, 2016.

[17] J. Chen, L. Song, M. Wainwright, and M. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892, 2018.

[18] C. Croux, J. Jagtiani, T. Korivi, and M. Vulanovic. Important factors determining fintech loan default: Evidence from a lendingclub consumer platform. *Journal of Economic Behavior & Organization*, 173:270–296, 2020.

[19] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617. IEEE, 2016.

[20] N. J. Diederich, C. G. Moore, S. E. Leurgans, T. A. Chmura, and C. G. Goetz. Parkinson disease with old-age onset: a comparative study with subjects with middle-age onset. *Archives of Neurology*, 60(4):529–533, 2003.

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

[22] V. Dunjko and H. J. Briegel. Machine learning & artificial intelligence in the quantum domain: A review of recent progress. *Reports on Progress in Physics*, 81(7):074001, 2018.

[23] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.

[24] N. Finkelstein and I. Shpitser. Deriving bounds and inequality constraints using logicalrelations among counterfactuals. In *Proceedings of the Thirty Sixth Conference on Uncertainty in Artificial Intelligence (UAI-20)*. AUAI Press, 2020.

[25] C. Fong, C. Hazlett, K. Imai, et al. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.

[26] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[27] N. Frosst and G. Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.

[28] Y. Gil, M. Greaves, J. Hendler, and H. Hirsh. Amplify scientific discovery with artificial intelligence. *Science*, 346(6206):171–172, 2014.

[29] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[30] N. Griefer. Weightit: Weighting for covariate balance in observational studies, 2019.

[31] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42, 2018.

[32] D. Gunning. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[33] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4):889–911, 2005.

[34] M. Hernan and J. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

[35] K. Hirano and G. W. Imbens. The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, pages 73–84. Wiley Blackwell, 2005.

[36] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[37] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, pages 1171–1220, 2008.

[38] M. C. Horowitz, G. C. Allen, E. Saravalle, A. Cho, K. Frederick, and P. Scharre. Artificial intelligence and international security. *Washington: Center for a New American Security (CNAS)*, 2018.

[39] W. Huan, Y. Wu, L. Zhang, and X. Wu. Fairness through equality of effort. In *Companion Proceedings of the Web Conference 2020*, pages 743–751, 2020.

[40] K. Imai and D. A. Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

[41] J. Jagtiani and C. Lemieux. The roles of alternative data and machine learning in fintech lending: Evidence from the lendingclub consumer platform. *Financial Management*, 48(4):1009–1029, 2019.

[42] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.

[43] R. Kass, T. Finin, et al. The need for user models in generating expert system explanations. *International Journal of Expert Systems*, 1(4), 1988.

[44] A. Khademi and V. Honavar. Algorithmic bias in recidivism prediction: A causal perspective. *arXiv preprint arXiv:1911.10640*, 2019.

[45] A. Khademi, S. Lee, D. Foley, and V. Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pages 2907–2914, 2019.

[46] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.

[47] P. Kitcher and W. Salmon. Van fraassen on explanation. *The Journal of Philosophy*, 84(6):315–330, 1987.

[48] N. Kreif and K. DiazOrdaz. Machine learning in policy evaluation: New tools for causal inference. In *Oxford Research Encyclopedia of Economics and Finance*. 2019.

[49] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler. Problems with shapley-value-based explanations as feature importance measures. *arXiv preprint arXiv:2002.11097*, 2020.

[50] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[51] A. Larrabee Sønderlund, E. Hughes, and J. Smith. The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, 50(5):2594–2618, 2019.

[52] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[53] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[54] M. K. Leung, A. Delong, B. Alipanahi, and B. J. Frey. Machine learning in genomic medicine: A review of computational problems and data sets. *Proceedings of the IEEE*, 104(1):176–197, 2015.

[55] S. Lipovetsky and M. Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

[56] Z. C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[57] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[58] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[59] C. F. Manski. *Identification for prediction and decision*. Harvard University Press, 2009.

[60] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[61] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley. Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, 2018.

[62] B. Mittelstadt, C. Russell, and S. Wachter. Explaining explanations in ai. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 279–288. ACM, 2019.

[63] E. Mjolsness and D. DeCoste. Machine learning for science: State of the art and future prospects. *Science*, 293(5537):2051–2055, 2001.

[64] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[65] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *arXiv preprint arXiv:1902.01876*, 2019.

[66] R. Nabi and I. Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

[67] B. Norgeot, B. S. Glicksberg, and A. J. Butte. A call for deep-learning healthcare. *Nature Medicine*, 25(1):14–15, 2019.

[68] J. Pearl. *Causality*. Cambridge university press, 2009.

[69] J. Pearl. The foundations of causal inference. *Sociological Methodology*, 40(1):75–149, 2010.

[70] R. Pelzer. Policing of terrorism using data from social media. *European Journal for Security Research*, 3(2):163–179, 2018.

[71] R. Pirracchio, M. L. Petersen, and M. van der Laan. Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, 181(2):108–119, 2015.

[72] A. Rajkomar, J. Dean, and I. Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[73] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[74] J. M. Robins, M. Á. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[75] C. Romero and S. Ventura. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[76] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.

[77] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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

[79] W. C. Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984.

[80] W. C. Salmon. *Causality and explanation*. Oxford University Press, 1998.

[81] P. Schwab, L. Linhardt, and W. Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

[82] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[83] R. Setiono and H. Liu. Understanding neural networks via rule extraction. In *IJCAI*, volume 1, pages 480–485, 1995.

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

[85] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017.

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

[87] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[88] J. Sirignano, A. Sadhwani, and K. Giesecke. Deep learning for mortgage risk. *arXiv preprint arXiv:1607.02470*, 2016.

[89] E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

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

[91] S. Thrun. Extracting rules from artificial neural networks with distributed representations. In *Advances in Neural Information Processing Systems*, pages 505–512, 1995.

[92] G. G. Towell and J. W. Shavlik. Extracting refined rules from knowledge-based neural networks. *Machine Learning*, 13(1):71–101, 1993.

[93] R. K. Tripathi, A. S. Jalal, and S. C. Agrawal. Suspicious human activity recognition: A review. *Artificial Intelligence Review*, 50(2):283–339, 2018.

[94] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, 2009.

[95] M. Tschannen, O. Bachem, and M. Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

[96] M. J. Van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[97] M. J. Van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[98] B. Van Fraassen. The pragmatic theory of explanation. *Theories of Explanation*, pages 135–155, 1988.

[99] M.-A. T. Vu, T. Adalı, D. Ba, G. Buzsáki, D. Carlson, K. Heller, C. Liston, C. Rudin, V. S. Sohal, A. S. Widge, et al. A shared vision for machine learning in neuroscience. *Journal of Neuroscience*, 38(7):1601–1607, 2018.

[100] F. Wang, L. P. Casalino, and D. Khullar. Deep learning in medicine—promise, progress, and challenges. *JAMA Internal Medicine*, 179(3):293–294, 2019.

[101] A. Waters and R. Miikkulainen. Grade: Machine learning support for graduate admissions. *AI Magazine*, 35(1):64–64, 2014.

[102] J. Wiens and E. S. Shenoy. Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*, 66(1):149–153, 2018.

[103] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[104] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[105] J. Zhang and E. Bareinboim. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, pages 3671–3681, 2018.

[106] Q. Zhao and T. Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, pages 1–10, 2019.

[107] Y. Zhu, D. L. Coffman, and D. Ghosh. A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference*, 3(1):25–40, 2015.

[108] J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.