CAUSAL INFERENCE VIA NONLINEAR VARIABLE DECORRELATION FOR HEALTHCARE APPLICATIONS

ABSTRACT

Causal inference and model interpretability research are gaining increasing attention, especially in the domains of healthcare and bioinformatics. Despite recent successes in this field, decorrelating features under nonlinear environments with human interpretable representations has not been adequately investigated. To address this issue, we introduce a novel method with a variable decorrelation regularizer to handle both linear and nonlinear confounding. Moreover, we employ association rules as new representations using association rule mining based on the original features to further proximate human decision patterns to increase model interpretability. Extensive experiments are conducted on four healthcare datasets (one synthetically generated and three real-world collections on different diseases). Quantitative results in comparison to baseline approaches on parameter estimation and causality computation indicate the model's superior performance. Furthermore, expert evaluation given by healthcare professionals validates the effectiveness and interpretability of the proposed model. Code will be publicly available after acceptance.

1 INTRODUCTION

With the rapid growth of Machine Learning (ML), healthcare ML research is becoming popular in the community. Such ML methods have shown encouraging capability for solving medically related problems, such as disease understanding, diagnosis, and treatment planning, by leveraging a large number of Electric Health Records (EHR). Although these methods bring benefits to both patients and healthcare professionals (Herpertz et al., 2017; Li et al., 2020), increasing concerns on judgment errors (Royce et al., 2019; Gandhi et al., 2006) as well as deficiency of understanding the workflow of ML systems (Croskerry, 2013) have become major road-blockers for future development and deployment of ML-based healthcare systems. An important factor behind this difficulty is that the designed black-box ML models are often associated with a limited capacity for performance analysis (Ahmad et al., 2018). Therefore, building interpretable ML models for healthcare becomes an imperative research direction.

To improve the interpretability of black-box models, more and more methods to enhance model interpretability are emerging (Du et al., 2019; Zafar & Khan, 2019). However, explanations of black-box models often cannot be perfectly faithful to the original models and leave out much information which cannot be made sense of (Rudin, 2019). In addition, traditional ML models might be influenced by the data they are trained on. In order to enhance the interpretability of the model and adapt it to human decision patterns, we introduce association rules in place of the original features.

Recently, most of the existing diagnostic algorithms focus on associative inference and are often not compatible with the situation caused by the incomplete distribution of datasets. Machine learning methods recognize diseases based on correlations and probability among patients' symptoms and medical history (Zhang et al., 2021; Kuang et al., 2020a), while doctors diagnose according to the best causal explanations corresponding to the symptoms (Imbens & Rubin, 2015). Recently, several methods have been proposed to address the agnostic distribution, including domain generalization which is becoming one of the most prominent learning paradigms (Muandet et al., 2013). Another school of research examines the distribution shift issue from a causal perspective, such as causal transfer learning (Rojas-Carulla et al., 2018) and Structural Causal Model (SCM) (Pearl, 2009) to identify causal variables based on the conditional independence test. In spite of their advantageous analytical qualities, these approaches are rarely employed in high-dimensional real-world applica-

tions due to the complex causal graph and strict assumptions. More recently, some researchers focus on more general methods under the stability guarantee by variable decorrelation through sample reweighting (Kuang et al., 2020b; Zhang et al., 2021; Kuang et al., 2018; 2021). They leveraged covariate balancing to eliminate the impact of confounding, assessing the effect of the target feature by reweighting the data so that the distribution of covariates is equalized across different target feature values. However, their model are **limited to linear environments or binary datasets**.

In this paper, we attempt to address the aforementioned difficulties by developing a novel method that is inherently more interpretable and can be applied to **nonlinear environments** for stable prediction. We utilize an association rule mining algorithm to extract rules as model features, thereby enhancing our model's interpretability. To enable our model to operate in nonlinear environments, we model the relationships between features with a $F(x)$ function, and perform the Taylor expansion on the $F(x)$ function. The second norm of the parameters from the first derivative to the last derivative are considered as our regularizer. Experiments conducted on both synthetic and real-world datasets demonstrate the efficacy of our approach. Promising results in improving the estimation of model parameters, and the stability of prediction over varying distributions in a nonlinear environment demonstrate the superior performance of the proposed method to previous methods.

The main contributions of our work are as follows:

(1) We expand the stable learning problem to a nonlinear environment under model misspecification and agnostic distribution so that stable learning can be widely applied in the real world;

(2) We combine machine learning with association rules to help domain specialists understand the model and enhance the interpretability of the model; and

(3) We demonstrate the superiority of our methods on synthetic and real-world datasets by calculating traditional metrics and causality. For medical datasets, we further invite specialized doctors to validate whether our model can produce the correct rules.

2 Related Work

2.1 Machine Learning Interpretability in Healthcare

Increasing efforts have been devoted to Machine Learning (ML) interpretability research to facilitate ML research and development of real-world applications, especially in healthcare. Among them, Generalized Additive Models (GAM) (Hastie & Tibshirani, 2017) are a set of classic methods with univariate terms providing straightforward interpretabilities. GA^{2}M-model (Lou et al., 2013) brings additional capability for real-world datasets with the selected interacting pairs based on GAMs. On the other hand, researchers focus on applying essentially interpretable models in healthcare domain. For example, Lee & Siau (2001) apply association rules to extract knowledge as complementary information for physicians' diagnosis. They also provide some strategies for patients based on the interpretation of association rules. Lately, Ahmed et al. (2021) apply association rules to detect major body organs in healthcare system. Sornalakshmi et al. (2021) reduce overhead communication when frequent data are extracted to improve association rules mining algorithm on healthcare datasets. However, the above models are still black-box models based on joint probability distribution without causal inference.

2.2 Association Rule Mining

Association rule mining is an important research direction that tries to identify interesting associations, frequent patterns, or causal structures (Perçın et al., 2019; Ordonez et al., 2006). In particular, association rules are able to discover predictive rules with numeric and categorical attributes. In diagnosis system, $X = x_1, x_2, ..., x_n$ represents the set of all symptoms. An association rule, noted as $X \Rightarrow Y$, indicates the disease Y is related to the symptoms X. Three metrics were proposed to evaluate the significance of rules: *support*$(X) = P(X)$ is the probability that the set appears in the total item set; *confidence*$(X \Rightarrow Y)$ = *support*$(X \cup Y)$/*support*(X) is a measure of reliability; *lift*$(X \Rightarrow Y)$ = *confidence*$(X \Rightarrow Y)$/*support*(Y) reflects the correlation between X and Y in the association rules (Bayardo Jr & Agrawal, 1999). In each rule, X is antecedent and Y is the consequent. The rules that satisfied the minimum support and confidence are called *strong* association

rules. Strong association rules, can also be divided into effective strong association rules and invalid strong association rules. How to extract strong association rules is an essential challenge. Apriori algorithm explores candidate-generation-and-test to obtain strong association rules (Borgelt & Kruse, 2002). Han proposed an effective method, the FP-Growth algorithm, to efficiently identity frequent patterns on large databases based on tree structures (Han et al., 2000). Yuan (2017) proposed an improved method based on the inherent defects of the Apriori algorithm by using a new mapping way and pruning frequent itemsets to improve efficiency. Association rules are interpretable models, whereas these methods always only consider extracting association rules based on connection instead of causality, and they do not combine rules and machine learning. In a diagnosis system, association rules are often inconsistent with the rules of doctors' diagnosis. Therefore, how to extract causal association rules that are consistent with doctors' diagnostic rules has become an important challenge.

2.3 CAUSAL INFERENCE

One key challenge in healthcare is the existence of both observed and unobserved confounders under different environments (Cui & Athey, 2022). Therefore, causal inference methods become popular for their natural fit to these problems. For example, causal inference methods with network and hierarchy structure allow researchers to ascribe causal explanations to data (Pearl, 2018; 2009). A completely constructed causal graph among various features based on an unconfoundedness assumption that helps to reduce the influence of confounders (Ma et al., 2021). In addition, a Differentiated Variable Decorrelation (DVD) algorithm is proposed to eliminate the correlations of various variables in different environments (Shen et al., 2020). Moreover, Xu et al. (2021) prove the effectiveness of stable learning and demonstrates the necessary of the stable prediction.

Stable Learning Given various environments $\mathbf{e} \in E$ within datasets $D^{\mathbf{e}} = (X^{\mathbf{e}}, Y^{\mathbf{e}})$, the task is to train a predictive model under the environment e_i which can achieve uniformly small error under the another environment e_j by learning the causality between features X^{e_i} and targets Y^{e_i}. Researchers propose the Deep Global Balancing Regression (DGBR) algorithm and Decorrelated Weighting Regression (DWR) algorithm for stable prediction across unknown environments. They successively regard each variable as a treatment variable by using a balancing regularizer with theoretical guarantee (Kuang et al., 2018; 2020b; Cui & Athey, 2022).

In Equation 1, W is sample weight, $\mathbf{X}_{\cdot,j}$ is the j^{th} variable in $\mathbf{X}$, and $\mathbf{X}_{\cdot,-j} = X/\{\mathbf{X}_{\cdot,j}\}$. With the global balancing regularizer in Equation 1, a Global Balancing Regression algorithm is proposed to optimize global sample weights and causality for classification task.

$$\begin{array}{ll} \min & \sum_{i=1}^{n} W_i \cdot \log\left(1 + \exp\left((1-2Y_i)\cdot(\mathbf{X}_i\beta)\right)\right) \\ \text{s.t.} & \sum_{j=1}^{p} \left\| \frac{\mathbf{x}_{-j}^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{i-j}^T (W \odot (1-\mathbf{X}_{\cdot,j}))}{W^T \cdot (1-\mathbf{X}_{\cdot,j})} \right\|_2^2 \leq \lambda_1 \end{array} \tag{1}$$

However, the above methods have some defects, making it difficult to deploy them on real world datasets. The regularizer of DGBR or DWR focuses on eliminating the linear confounding under various linear environments. In addition, such methods that forcibly delete mutual connections ignore information in the intersecting area. For example, as shown in Figure 1, forcibly eliminating the correlation may leave only three areas: A, B, C and ignore the other areas, where A, B, C can represent three kinds of features. In fact, the causal effect of the three features should be the union of these areas, whereas DGBR may result in the loss of mutual information.

3 METHOD

To achieve stable learning and estimation with unbiased treatment effect, we make three assumptions for our model:

(i) The set of strong causality rules is a subset of the set of rules with strong correlations;

(ii) There do not exist massive observed confounders that cause the diagnostic rules dependent on another unobserved confounding; and

(iii) Causal rules tend to include smaller antecedents.

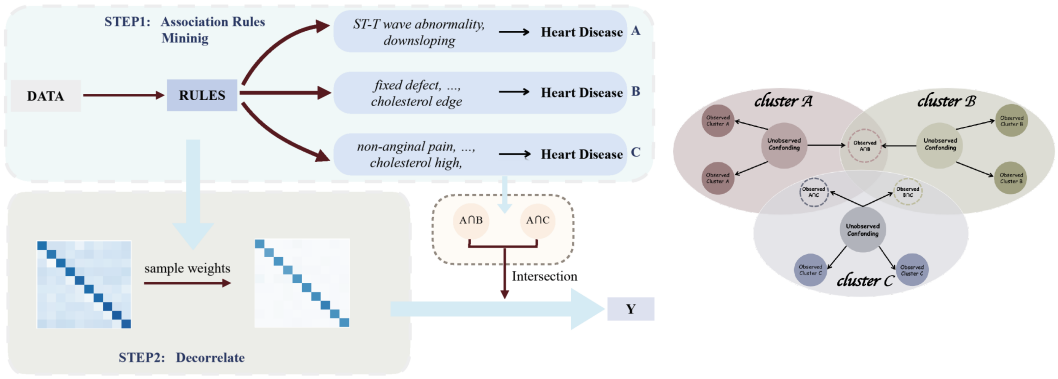

Figure 1: Causality consists of intersecting regions and disjoint regions. Our model will extract rules from the original datasets and feed them into decorrelation parts to calculate the causality with mutual information.

The first and second assumptions imply that we can extract all the diagnostic rules with strong correlation, and then assign weights to these rules based on the causal effect. The third assumption implies that diagnosis rules with high confidence often result in overfitting due to redundant antecedents. Therefore, we prefer to choose more robust rules with shorter antecedents. In this paper, we propose an interpretable model based on association rules and causal inference for EHR datasets to obtain the causality between features through a three-stage process:

A. Mining Association Rules and Transformation Rules: First, we adopt the Apriori algorithm (Agrawal et al., 1994) to obtain association rules and construct a rule matrix for positive and negative samples respectively to avoid asymmetrically distributed data. Rule representations $< \mathscr{A}_i, \mathscr{C}_i | \theta_i >$ are then constructed, where $\mathscr{A}_i$ is the antecedent of the rule R_i, $\mathscr{C}_i$ is the consequent of the rule and θ_i is the confidence of the rule, frequent function is to calculate the confidence of each rule:

$$\{R\} = \{\cup_i R_i\} = \{\cup_i frequent(\mathscr{A}_i \cup \mathscr{C}_i)\}$$

According to the rules generated, rule sets $\cup_i \{X_i :< \mathscr{A}_i, \mathscr{C}_i | \theta_i >\}$ are built for dataset D where each rule is considered as a feature. Rule sets are then transformed into one-zero matrix X leveraging one-hot encoding.

B. Selecting Rules: Massive rules could be generated during the mining process, causing redundancy or even negative effects. To extract rules with strong correlations between features, we introduce an integer programming objective function:

$$\begin{array}{ll} \text{Min} & \|W\|_2^2 + \|\max(0, 1 - Yh(x)\|_2^2 \\ & h(x) = (W^T X \odot rep(\mathbb{I}(R > 0), n)\theta + b) \\ \text{s.t.} & \sum_i \mathbb{I}(R_i > 0) \leq \lambda_1 \\ & \sum_i \mathbb{I}(R_i > 0) \geq \lambda_2 \\ & \{R_i\} \in \{0, 1\} \end{array} \tag{2}$$

where $\odot$ refers to the Hadamard product and $\mathbb{I}(R > 0)$ is the indicator function converting R, a set of rules, to a one-zero vector with $1 * r$ dimension. The value of the indicator function equals to one when the frequency of the rule is more than zero, otherwise it equals to zero. The estimated parameters of $rep(\mathbb{I}(R > 0), n)$ function are W, and b is the estimated bias. The function is defined to expand the vector $\mathbb{I}(R^{1*r} > 0)$ to a matrix with dimension $n * r$ where all rows are the same as the first row. λ_1 and λ_2 represent the bonds for the number of the selected rules.

Since we only consider a binary-classification problem here, which is a common setting for most healthcare diagnosis problems, we take the inverse of the confidence of the negative class rule as the score. However, the number of rules mined by the association rule algorithm, e.g. Apriori, could be large, resulting an extremely high dimension of R to be able to fit in Equation 2. Therefore, we propose to delete one redundant rule at a time during each n-fold cross-validation run based on a

feature ranking criteria w_i^2. Details of rule selection can be found in Algorithm 1 in the Appendix under *RulesSelection*.

Although redundant rules are removed accordingly, redundant items in rules could still impact the performance of the model. In addition, redundant items in different rules could be easily replaced with other rules. To solve this problem, we perform an iterative process to delete one item of each rule at a time which brings an updated R with reduced dimension. Then we can reconstruct cross-validation sets and feed data into SVM models to get an average accuracy. The item that improves model's average accuracy the most will be deleted at every iteration. This step is summarized in Algorithm 1 in the Appendix under *ItemReduce*.

C. Computing Causality Relationship: To better handle real-world nonlinear relationships, we model nonlinear relationships under Taylor expansion with a function $\mathcal{F}(x)$ as is shown in Equation 3. Each fixed point's derivatives can be considered as parameters to be solved by converting into a polynomial fitting problem due to the condition that two polynomials are equal only when both their degree and coefficients are the same.

$$\begin{aligned} x_{p_1} \sim \quad & \mathcal{F}(x_j) = f_{p_1p_2}(x_{p_2}(0)) + f'_{p_1p_2}(x_{p_2}(0))x_{p_2} + \\ & \frac{f''_{p_1p_2}(x_{p_2}(0))}{2!}x_{p_2}^2 + \ldots + \frac{f^{(p)}_{p_1p_2}(x_{p_2}(0))}{p!}x_{p_2}^p + R_p(x_{p_2}) \end{aligned} \tag{3}$$

where $x_{p_1}(0)$ and $x_{p_2}(0)$ are two different features which are expanded at 0 by using Taylor expansion. The elimination of the impact of intersecting areas is achieved by balancing the weight W as is represented in Equation 4. If x_{p_1} and x_{p_2} are independent and nonlinearly uncorrelated, the derivatives of their relation functions are all 0: $\|\{\mathscr{F}_{p_2\to p_1}\} / \{f_{p_1p_2}(x_{p_2}(0))\}\| = 0$ where $\mathscr{F}_{p_2\to p_1}$ are the relationship function between $x_{p_1}(0)$ and $x_{p_2}(0)$ can be calculated using Equation 5

$$\begin{aligned} & \min_{\mathscr{F}_{p_2\to p_1}} R_p(x)^2 \equiv \textstyle\sum_{p_1\neq p_2}\sum_{i=1}^{n}\left[w_i x_{ip_2} - \mathcal{F}\left(w_i x_{ip_1}\right)\right]^2 \\ & \Rightarrow \mathscr{X}_{p_2}\left(w_i x_{p_2}\right)\mathscr{F}_{p_2\to p_1} = \mathscr{Y}_{p_1} \end{aligned} \tag{4}$$

$$\mathscr{X}_{p_2} = \begin{bmatrix} n & \sum_{i=1}^n w_i x_{ip_2} & \cdots & \sum_{i=1}^n w_i^k x_{ip_2}^k \\ \sum_{i=1}^n w_i x_{ip_2} & \sum_{i=1}^n w_i^2 x_{ip_2}^2 & \cdots & \sum_{i=1}^n w_i^{k+1} x_{ip_2}^{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n w_i^k x_{ip_2}^k & \sum_{i=1}^n w_i^{k+1} x_{ip_2}^{k+1} & \cdots & \sum_{i=1}^n w_i^{2k} x_{ip_2}^{2k} \end{bmatrix}$$

$$\mathscr{F}_{p_2\to p_1} = \begin{bmatrix} f_{p_1p_2}(x_{p_2}(0)) \\ f'_{p_1p_2}(x_{p_2}(0)) \\ \vdots \\ f^{(p)}_{p_1p_2}(x_{p_2}(0)) \end{bmatrix}, \mathscr{Y}_{p_1} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{j=1}^n x_i y_i \\ \vdots \\ \sum_{i=1}^n x_i^k y_i \end{bmatrix}$$

$$\mathscr{F}_{p_2\to p_1} = \left(\mathscr{X}_{p_2}^{\mathrm{T}}\mathscr{X}_{p_2}\right)^{-1}\mathscr{X}_{p_2}^{\mathrm{T}}\mathscr{Y}_{p_1} \tag{5}$$

where w_i is the weight for each sample and n is the number of datasets. Combined with Figure 1, physical meaning can be given to the above variables: $f(\theta)$ is the correlation between each feature and target; C is the factor to expand the influence of the intersection area to get the real causality comparing with W applied to eliminate the influence of the public area:

$$\begin{aligned} \text{Min} \quad & \tfrac{1}{2}\beta^T\beta + \textstyle\sum_{i=1}^n (W_i + C)\max\left(0, 1 - y_i\left(\beta^T\phi\left(x_i\right) + b\right)\right) \\ & \|\mathscr{F}^{(i)}_{p_2\to p_1, i>0}\|_2^2 \leq \gamma, \|W\|_2^2 \leq \lambda_1, \left(\textstyle\sum_{k=1}^n W_k - 1\right)^2 \leq \lambda_2 \end{aligned} \tag{6}$$

When we have a smaller γ value, the difference between β and the true correlation coefficient (disjoint region and the target) will become smaller, resulting greater mutual information loss.

Lemma 1. *If the number of features in the datasets and the terms in the Taylor expansion are fixed, when $n \to \infty$ there exists $W \succeq 0$ such that*

$$\lim_{n\to\infty} \|\mathscr{F}^{(i)}_{p_2\to p_1, i>0}\|_2^2$$

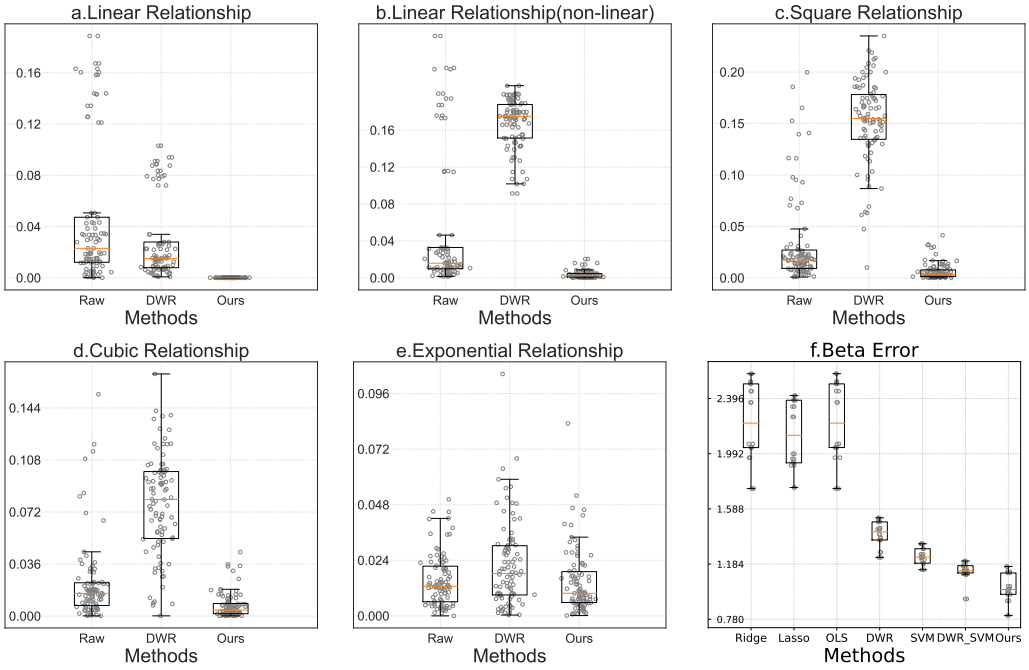

Figure 2: Figures (a)-(d) describe the distribution of the Pearson Coefficient values among various relationships. Figure (a) reports the β errors of different models. Figure (f) is under a linear environment and other figures are under nonlinear environments. Our model is able to provide the greatest reduction of both linear and nonlinear relationships.

4 Experiment

4.1 Validation on A Synthetic Dataset

To examine the proposed constraints' effect on eliminating linear and nonlinear connotation relationships, we follow previous work (Kuang et al., 2020b) to conduct evaluations on synthetically generated datasets. The details of experiment settings and baseline methods can be found in the Appendix (A.2). Notice that a different objective function 7 is built for regression task, where W_i is the sample weight and the variable ζ is slack variable. In this experiment, we only expand two terms by the Taylor expansion.

$$\begin{aligned} \min_{w,b,\zeta,\zeta^*} \frac{1}{2} w^T w + \sum_{i=1}^{n} (C + W_i)(\zeta_i + \zeta_i^*) \\ \text{subject to } y_i - w^T \phi(x_i) - b \leq \varepsilon + \zeta_i \\ w^T \phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \geq 0, i = 1, \ldots, n \\ \|\mathscr{F}^{(i)}_{p_2 \to p_1, i>0}\|_2^2 \leq \gamma \\ \|W\|_2^2 \leq \lambda_1, (\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_2 \end{aligned} \tag{7}$$

4.1.1 Results

To compare two kinds of regularizers, we apply Pearson Correlation to calculate the relationship strength among features. Since Pearson Correlation can only describe linear relationship, we construct nonlinear pairs $\mathbf{WV_i}$ with $(\mathbf{WV})^{\mathbf{2}}_{\mathbf{j}}$, $(\mathbf{WV})^{\mathbf{3}}_{\mathbf{j}}$ and $exp(\mathbf{WV}_j)$ in addition to $\mathbf{WV_i}$ with $\mathbf{WV_j}$. The result can be found in Figure 2. Both DWR and the proposed regularizer can handle pure linear relationships (experimental environment(A)) but some improvement is achieved from the proposed regularizer. As we add nonlinear relationships to the linear experimental environment, DWR start to have difficulty with the linear relationship part while the proposed method is still able to reduce a large amount of the relationships. For nonlinear environments, compared with the

Table 1: Results under varying sample size *n* and number of variables within nonlinear environments.

	n=1000, m=5			n=1000, m=10			n=1000, m=15		
	β_S Error	β_V Error	β Error	β_S Error	β_V Error	β Error	β_S Error	β_V Error	β Error
OLS	3.357	0.430	1.894	3.605	0.729	2.167	3.823	0.866	2.345
Lasso	3.390	0.326	1.858	3.586	0.647	2.117	3.940	0.390	2.165
Ridge	3.357	0.430	1.893	3.604	0.729	2.166	3.822	0.866	2.344
SVM	2.067	0.240	1.153	2.273	0.375	1.324	2.366	0.410	1.388
DWR	2.279	0.249	1.264	2.566	0.658	1.612	3.258	1.182	2.220
DWR_SVM	1.799	0.303	1.051	2.077	0.483	1.280	2.494	0.918	1.706
OUR	**1.555**	**0.199**	**0.877**	**1.898**	**0.373**	**1.135**	**2.265**	**0.382**	**1.323**
	n=2000, m=5			n=2000, m=10			n=2000, m=15		
	β_S Error	β_V Error	β Error	β_S Error	β_V Error	β Error	β_S Error	β_V Error	β Error
OLS	3.253	0.444	1.849	3.521	0.630	2.075	4.071	0.561	2.316
Lasso	3.278	0.250	1.764	3.490	0.473	1.982	4.260	0.168	2.214
Ridge	3.253	0.444	1.848	3.520	0.630	2.075	4.071	0.561	2.316
DWR	2.147	0.231	1.189	2.244	0.493	1.369	2.749	0.974	1.861
SVM	2.020	0.271	1.145	2.158	0.315	1.237	2.453	0.349	1.401
DWR_SVM	1.675	0.305	0.990	1.861	0.407	1.134	2.317	0.572	1.445
OUR	**1.544**	**0.214**	**0.879**	**1.719**	**0.292**	**1.006**	**2.125**	**0.323**	**1.224**
	n=3000, m=5			n=3000, m=10			n=3000, m=15		
	β_S Error	β_V Error	β Error	β_S Error	β_V Error	β Error	β_S Error	β_V Error	β Error
OLS	3.297	0.335	1.816	3.593	0.579	2.086	3.736	0.611	2.173
Lasso	3.279	0.074	1.677	3.803	0.179	1.991	3.703	0.527	2.115
Ridge	3.297	0.335	1.816	3.593	0.579	2.086	3.735	0.611	2.173
DWR	2.178	0.150	1.164	1.970	0.415	1.192	2.610	0.547	1.578
SVM	2.066	0.217	1.141	2.046	0.338	1.192	2.261	0.329	1.295
DWR_SVM	1.764	0.284	1.024	1.833	0.312	1.072	2.082	0.484	1.283
OUR	**1.748**	**0.065**	**0.907**	**1.618**	**0.171**	**0.894**	**2.007**	**0.325**	**1.166**

original unweighted dataset, DWR unexpectedly increases nonlinear relationships where there are no existing nonlinear relationships (square, cubic and exponential). Instead, our model can deal with nonlinear relationships and reduce nonlinear relationships. To test our algorithm perform the best on various sample sizes and the number of features, we calcualte the β_S and $beta_V$ errors: $error(\beta) = \sum_i |\beta_{true} - \beta|$, based on 9 kinds of datasets 1.

We conduct a series of ablative studies to evaluate the stability of our model. Table 2 summarizes the results of the experiment with various values of C and Lagrange penalty operators γ, and λ. For each sell, we fix the Lagrange penalty operators and increase C to calculate the β errors and RMSE errors. The higher C indicates higher integrated mutual information is fed into the model and magnifies the impact of confounding. Higher γ values will reduce more confounding effects and diminish mutual information. Here we choose the best parameters ($\gamma = 600, \lambda = 0.0005, C = 0.5$) based on the smallest RMSE also with a smaller β error comparing to ($\gamma = 1000, \lambda = 0.0005, C = 0.5$) which has the same RMSE.

To further confirm that the coefficients estimated by our model are based on causality, we repeat experiments 50 times to calculate $\sum \|\beta - \hat{\beta}\|$, where β and $\hat{\beta}$ represent the true value and estimated parameters, respectively. In Figure 2, we find that the difference between the estimated parameters and the true values is smaller with our model, compared to other models in the nonlinear environment. Notice that our model achieves much smaller distribution variance as well as much smaller average values of β errors comparing to baselines. Although the regularizer of DWR can solve the stable problem in linear environments, it retains or expands nonlinear confounding in the nonlinear environments. From the above results, we find that our model is able to reduce correlations among all predictors and avoid being affected by nonlinear confounding, resulting a reduced estimation bias in more general environments.

4.2 Validation on Three Real-World Datasets

To further validate the effectiveness of our model in real-world scenarios, we perform experiments on three different EHR datasets. All data are prepossessed to ensure no sensitive information is exposed.

Table 2: Ablative study. γ and λ are Lagrange penalty operators.

	$\gamma = 600, \lambda = 0.0001$			$\gamma = 600, \lambda = 0.0005$			$\gamma = 600, \lambda = 0.001$		
	$C = 0$	$C = 0.5$	$C = 1$	$C = 0$	$C = 0.5$	$C = 1$	$C = 0$	$C = 0.5$	$C = 1$
β_S Error	1.956	1.919	1.996	1.769	**1.926**	2.003	1.956	2.026	2.073
β_V Error	0.238	0.179	0.166	0.245	0.187	0.178	0.246	0.199	0.175
$RMSE$ Error	4.943	4.732	4.680	4.854	**4.726**	4.675	4.951	4.856	4.808
	$\gamma = 800, \lambda = 0.0001$			$\gamma = 800, \lambda = 0.0005$			$\gamma = 800, \lambda = 0.001$		
	$C = 0$	$C = 0.5$	$C = 1$	$C = 0$	$C = 0.5$	$C = 1$	$C = 0$	$C = 0.5$	$C = 1$
β_S Error	1.954	2.022	2.070	1.784	2.025	2.068	1.960	2.019	2.009
β_V Error	0.240	0.197	0.172	0.234	0.195	0.176	0.245	0.195	0.174
$RMSE$ Error	4.945	4.859	4.825	4.849	4.860	4.793	4.961	4.858	4.674
	$\gamma = 1000, \lambda = 0.0001$			$\gamma = 1000, \lambda = 0.0005$			$\gamma = 1000, \lambda = 0.001$		
	$C = 0$	$C = 0.5$	$C = 1$	$C = 0$	$C = 0.5$	$C = 1$	$C = 0$	$C = 0.5$	$C = 1$
β_S Error	1.962	2.022	2.075	1.959	1.928	2.073	1.962	2.024	2.006
β_V Error	0.242	0.196	0.173	0.250	0.187	0.178	0.244	0.189	0.169
$RMSE$ Error	4.938	4.859	4.812	4.950	**4.726**	4.811	4.947	4.854	4.672

4.2.1 DATASETS AND SETTINGS

Heart Disease is retrieved from the repository of the University of California, Irvine (Asuncion & Newman, 2007). We follow previous work to use 13 of 76 attributes: *Age*, *Sex*, *cp*, *threstbps*, *chol*, *fbs*, *restecg*, *thalach*, *exang*, *oldpeak*, *slope*, *cam* and *thal*.

Esophageal Cancer consists of data from 261 patients who underwent esophagectomy for esophageal cancer between 2009 and 2018. The collected characteristics include patient demographics, medical and surgical history, clinical tumor staging, adjuvant chemoradiotherapy, esophagectomy procedure type, postoperative pathologic tumor staging, adjuvant chemoradiotherapy, postoperative complications, cancer recurrence, and mortality.

Cauda Equina Syndrome (CES) is extracted from the Statewide Planning and Research Cooperative System (SPARCS) (of Health & Bureau, 1984), a comprehensive database of all payers for all hospitalizations in New York State Joo et al. (2022).

Based on diagnostic and procedure codes, patients with CES who underwent surgery between 2000 and 2015 were selected. Patient demographics (age, gender, race, comorbidities, and insurance status) and hospital characteristics (measured by hospital bed number quartiles).

Pre-Processing: We convert the continuous variables into categorical variables before feeding them to the model. To handle missing data in the datasets, we adopted MICE (Multiple imputations by chained equations) by transforming imputation problems into estimation problems where each variable will be regressed on the other variables. This method provides promising flexibility since every variable can be assigned a suitable distribution (Wulff & Jeppesen, 2017). Then we apply the SMOTE algorithm (Fernández et al., 2018) to address the class imbalance issue in our datasets.

Feature Selection: Redundant information in EHR datasets may cause noise and irrelevant information during feature extraction. A feature selection method (Guyon et al., 2002) is adopted. To improve the robustness of the model, we divide the dataset randomly into five groups for cross-validation. Each time we extract one group as the test set to analyze and measure the average performance in the feature selection process. Due to the high complexity of our model, we apply and compare the four baseline models: XGboost, SVM, Logistic Regression, and Random Forest to extract important features in feature selection and input the set of the features with the highest average AUROC scores into our model. In the end, we extract 13, 47, and 45 features for *Heart Disease*, *Esophageal Cancer*, and *Cauda Equina Syndrome*, respectively.

Comparison to Baseline Models: After the feature selection process, we transform continuous features into categorical variables before inputting data into our model and then applying one-hot encoding to convert categorical attributes into numeric, since the association rule mining algorithm in our paper cannot accept continuous features. However, the data set, filtered by feature selection, is directly fed into baseline models since forcing the continuous features to be discretized may lead to worse performance of the model. We assign 20% data into test datasets and compare our model

Table 3: Prediction performances over various healthcare datasets.

	Non Rule-based					Rule-based						
	XGBoost	RF	SVM	LR	MLP	XGBoost	RF	SVM	LR	MLP	DWR	Ours
Heart Disease												
Accuracy	0.903	0.887	0.885	0.869	0.947	0.869	0.868	**0.960**	0.934	0.878	0.937	**0.960**
F1	0.880	0.863	0.899	0.882	0.952	0.882	0.879	0.963	0.940	0.892	0.943	**0.964**
Precision	0.880	0.846	0.886	0.882	0.941	0.857	0.864	**0.972**	0.939	0.850	0.931	0.966
Recall	0.880	0.880	0.912	0.882	0.963	0.909	0.897	0.956	0.945	0.940	0.958	**0.964**
Causality	-	-	-	-	-	0.398	0.274	0.455	0.458	0.402	0.320	**0.528**
Esophageal Cancer												
Accuracy	0.788	0.750	0.827	0.808	0.750	0.738	0.727	**0.900**	0.812	0.846	0.854	**0.900**
F1	0.776	0.683	0.809	0.800	0.735	0.708	0.697	**0.888**	0.783	0.824	0.825	0.885
Precision	0.704	0.737	0.827	0.808	0.720	0.723	0.692	0.867	0.804	0.843	0.842	**0.874**
Recall	0.864	0.636	0.792	0.833	0.750	0.699	0.713	**0.913**	0.771	0.812	0.811	0.900
Causality	-	-	-	-	-	0.130	0.236	0.281	0.327	0.160	0.314	**0.339**
Cauda Equina Syndrome												
Accuracy	0.788	0.75	0.827	0.808	0.750	0.883	0.779	0.887	0.886	0.891	0.891	**0.893**
F1	0.776	0.683	0.809	0.800	0.735	0.880	0.780	0.883	0.882	**0.888**	0.887	**0.888**
Precision	0.704	0.737	0.827	0.808	0.720	0.818	0.706	0.825	0.822	0.831	0.831	**0.834**
Recall	0.864	0.636	0.792	0.833	0.750	0.951	0.874	0.950	0.952	**0.953**	**0.953**	0.951
Causality	-	-	-	-	-	0.231	0.298	0.279	0.132	0.262	0.308	**0.477**

with five baseline models: Logistic Regression, Random Forest, XGboost, SVM and MLP as shown in Appendix A.3

4.2.2 RESULTS

To measure the performance of models, we calculate accuracy, precision, recall and F1 scores. The result is shown in Appendix A.3, Table 3. In addition to calculating the metrics of the traditional models, we input the filtered rules as one-zero matrix X into the baselines rather than the original datasets. In *Heart Disease* and *Esophageal Cancer* datasets, rules do help XGBoost, Random Forest and MLP to improve the performance, while in *Cauda Equina Syndrome* datasets rules can improve the performance of all models. For SVM and Logistic Regression, the effect of the model can be greatly improved after combining the rules. Our model generally performs the best on all three datasets, similar to the SVM performance, while achieving high interpretability as discussed below.

Combining the experiments in the previous section, better performance is not equivalent to obtaining the real rules. To compare the causality calculated by our model and baselines, we ask three groups of doctors of the corresponding domains to score each rule. Three groups of doctors are from Cardiology, ENT and Neurosurgery departments, and each group consists of three doctors. we apply the models to calculate the importance of features to score each rule. To verify rating consistency between our model and doctors leveraging Spearman Coefficients. Results can be found in Table 3. As can be observed, causality rankings of the baseline models vary greatly, indicating unstable performance. In these datasets, the causal value of our model is higher than other baselines, implying that the scoring of our model is more consistent with the standard of doctors.

5 CONCLUSION

In this paper, we present a causal inference approach focusing on interpretability and nonlinear environments for healthcare applications. The proposed method extracts association rules from the raw features as new representations to be used by our model. A novel regularizer that is capable for handling both linear and nonlinear confoundings is constructed to enable our model's adaption to real-world applications. The superior performances on a synthetic dataset and three real-world EHR datasets from different domains compared to baseline methods validate both the effectiveness and generalizability of the proposed method. Consistent ratings with healthcare professionals on the extracted rules further validate the model's interpretability, while not sacrificing accuracy.

REFERENCES

S Agatonovic-Kustrin and Rosemary Beresford. Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, 22(5):717–727, 2000.

Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pp. 487–499. Citeseer, 1994.

Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pp. 559–560, 2018.

Imran Ahmed, Gwanggil Jeon, and Francesco Piccialli. A deep-learning-based smart healthcare system for patient's discomfort detection at the edge of internet of things. *IEEE Internet of Things Journal*, 8(13):10318–10326, 2021.

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

Roberto J Bayardo Jr and Rakesh Agrawal. Mining the most interesting rules. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 145–154, 1999.

Raghu Bollapragada, Jorge Nocedal, Dheevatsa Mudigere, Hao-Jun Shi, and Ping Tak Peter Tang. A progressive batching l-bfgs method for machine learning. In *International Conference on Machine Learning*, pp. 620–629. PMLR, 2018.

Christian Borgelt and Rudolf Kruse. Induction of association rules: Apriori implementation. In *Compstat*, pp. 395–400. Springer, 2002.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Pat Croskerry. From mindless to mindful practice—cognitive bias and clinical decision making. *N Engl J Med*, 368(26):2445–2448, 2013.

Peng Cui and Susan Athey. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4(2):110–115, 2022.

Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.

Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

Tejal K Gandhi, Allen Kachalia, Eric J Thomas, Ann Louise Puopolo, Catherine Yoon, Troyen A Brennan, and David M Studdert. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. *Annals of internal medicine*, 145(7):488–496, 2006.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12, 2000.

Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Routledge, 2017.

Sabine C Herpertz, Steven K Huprich, Martin Bohus, Andrew Chanen, Marianne Goodman, Lars Mehlum, Paul Moran, Giles Newton-Howes, Lori Scott, and Carla Sharp. The challenge of transforming the diagnostic system of personality disorders. *Journal of personality disorders*, 31(5): 577–589, 2017.

Arthur E Hoerl and Robert W Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

Graeme D Hutcheson. Ordinary least-squares regression. *L. Moutinho and GD Hutcheson, The SAGE dictionary of quantitative management research*, pp. 224–228, 2011.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Peter Joo, Weijian Li, Amy Phan, Gabriel Ramirez, Caroline P Thirukumaran, Jiebo Luo, Emmanuel N Menga, and Addisu Mesfin. 96. health care disparities in complication and mortality rates following surgical management of cauda equina syndrome. *The Spine Journal*, 22(9):S52–S53, 2022.

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1617–1626, 2018.

Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, and Zhichao Jiang. Causal inference. *Engineering*, 6(3):253–263, 2020a.

Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. Stable prediction with model misspecification and agnostic distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4485–4492, 2020b.

Kun Kuang, Hengtao Zhang, Runze Wu, Fei Wu, Yueting Zhuang, and Aijun Zhang. Balance-subsampled stable prediction across unknown test data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(3):1–21, 2021.

Sang Jun Lee and Keng Siau. A review of data mining techniques. *Industrial Management & Data Systems*, 2001.

Weijian Li, Wei Zhu, E Ray Dorsey, and Jiebo Luo. Predicting parkinson's disease with multimodal irregularly collected longitudinal smartphone data. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 1106–1111. IEEE, 2020.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 623–631, 2013.

Jing Ma, Yushun Dong, Zheng Huang, Daniel Mietchen, and Jundong Li. Assessing the causal impact of covid-19 related policies on outbreak dynamics: A case study in the us. *arXiv preprint arXiv:2106.01315*, 2021.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

New York (State). Department of Health and New York (State). SPARCS Bureau. *Statewide Planning and Research Cooperative System Annual Report Series*. New York State Department of Health, 1984.

Carlos Ordonez, Norberto Ezquerra, and Cesar A Santana. Constraining and summarizing association rules in medical data. *Knowledge and information systems*, 9(3):1–2, 2006.

Mahesh Pal. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222, 2005.

Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

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

İbrahim Perçın, Fatma Hilal Yağin, Emek Güldoğan, and Saim Yoloğlu. Arm: An interactive web software for association rules mining and an application in medicine. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–5. IEEE, 2019.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Celeste S Royce, Margaret M Hayes, and Richard M Schwartzstein. Teaching critical thinking: a case for instruction in cognitive biases to reduce diagnostic errors and improve patient safety. *Academic Medicine*, 94(2):187–194, 2019.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Zheyan Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bo Li, and Zhitang Chen. Stable learning via differentiated variable decorrelation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2185–2193, 2020.

M Sornalakshmi, S Balamurali, M Venkatesulu, M Navaneetha Krishnan, Lakshmana Kumar Ramasamy, Seifedine Kadry, and Sangsoon Lim. An efficient apriori algorithm for frequent pattern mining using mapreduce in healthcare data. *Bulletin of Electrical Engineering and Informatics*, 10(1):390–403, 2021.

Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Jesper N Wulff and Linda Ejlskov Jeppesen. Multiple imputation by chained equations in praxis: guidelines and review. *Electronic Journal of Business Research Methods*, 15(1):41–56, 2017.

Renzhe Xu, Peng Cui, Zheyan Shen, Xingxuan Zhang, and Tong Zhang. Why stable learning works? a theory of covariate shift generalization. *arXiv preprint arXiv:2111.02355*, 2021.

Xiuli Yuan. An improved apriori algorithm for mining association rules. In *AIP conference proceedings*, volume 1820, pp. 080005. AIP Publishing LLC, 2017.

Muhammad Rehman Zafar and Naimul Mefraz Khan. Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. *arXiv preprint arXiv:1906.10263*, 2019.

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5372–5382, 2021.

A APPENDIX

A.1 ALGORITHM

We combine algorithm 1 with object function 2 to select the robust rules and prune the redundant items. In the ***RulesSelection*** function, we delete one rule each time with lowest $\|w\|_2^2$ and save the rule sets with the highest accuracy. In the ***ItemReduce*** function, we apply cross-validation to train SVM model and save the item sets with best accuracy.

Algorithm 1 Rules Selection and Item Reduction

Input: $Rules\{X_i\}$ are the association rules obtained by Apriori algorithm with training datasets. $data$ is EHR datasets.
Output: $Bestrules$

function RULESSELECTION($Rules, data$)
  $Bestrules \leftarrow Rules$
  $Objfunction$ is objective function
  $Select \leftarrow Bestrules$
  $Bestaccuracy \leftarrow Select$
  $Lastrules \leftarrow \varnothing$
  while $Select \neq Lastrules$ **do**
    $Lastrules \leftarrow Select$
    $w \leftarrow \operatorname{argmin} Objectfunction(Select, data)$
    $Selected \leftarrow \operatorname{argmin} w_i^2$
    $Temprules \leftarrow \{Bestrules\}/\{Selected\}$
    $Tempaccuracy \leftarrow Temprules$
    if $Tempaccuracy > Bestaccuracy$ **then**
      $Bestaccuracy \leftarrow Tempaccuracy$
      $Select \leftarrow Temprules$
    end if
  end while
  return $Bestrules$
end function
function ITEMREDUCE($Bestrules, data$)
  $Bestauc \leftarrow SVM(Bestrules, data)$
  $Lastrules \leftarrow \varnothing$
  while $Bestrules \neq Lastrules$ **do**
    $Item \leftarrow \operatorname{argmax} SVM(\{Bestrules\}/\{Item\})$
    $Accuracy \leftarrow SVM(\{Bestrules\}/\{Item\})$
    if $Accuracy \geq Bestauc$ **then**
      $Bestauc \leftarrow Accuracy$
      $Bestrules \leftarrow \{Bestrules\}/\{Item\}$
    end if
  end while
  return $Bestrules$
end function

A.2 DATASETS AND SETTINGS

In addition to the linear settings, we propose to include nonlinear evaluations under a nonlinear environment:

Linear Environment: For this setting, we construct features $\mathbf{S}$ that causes unstable $\mathbf{V}$ by auxiliary variables z with linear relationship among features only.

$$\begin{aligned}\mathbf{Z}_{,1}, \cdots, \mathbf{Z}_{,p} &\overset{iid}{\sim} \mathcal{N}(0,1), \mathbf{X}_{,1}, \cdots, \mathbf{X}_{,p_v} \overset{iid}{\sim} \mathcal{N}(0,1) \\ \mathbf{S}_{,i} &= 0.8 * \mathbf{Z}_{,i} + 0.2 * \mathbf{Z}_{,i+1}, i = 1, 2, \cdots, p_s \\ \mathbf{V}_{\cdot,j} &= 0.8 * \mathbf{X}_{\cdot,j} + 0.2 * \mathbf{X}_{\cdot,j+1} + \mathcal{N}(0,1)\end{aligned}$$

Nonlinear Environment: In this setting, we combined square relationship and exponential relationship to generate various environment including potential nonlinear confounding to test our reweighted regularizer.

$$\begin{aligned}\mathbf{V}_{\cdot,j} &= \mathbf{X}_{\cdot,j} + 0.4 * \mathbf{X}_{\cdot,j+1} + 0.4 * exp(\mathbf{X}_{\cdot,j+1}) \\ &\quad + 0.4 * \mathbf{X}^2_{\cdot,j+1} + 0.1 * \mathbf{X}^3_{\cdot,j+1} + \mathcal{N}(0,1) \\ \mathbf{S}_{\cdot,j} &= \mathbf{Z}_{\cdot,j} + 0.4 * \mathbf{Z}_{\cdot,j+1} + 0.4 * exp(\mathbf{Z}_{\cdot,j+1}) \\ &\quad + 0.4 * \mathbf{Z}^2_{\cdot,j+1} + 0.1 * \mathbf{Z}^3_{\cdot,j+1} + \mathcal{N}(0,1)\end{aligned}$$

To further test the robustness of our algorithm, we assume that there are unobserved nonlinear terms, and construct the label Y as shown in Equation 8. Combined with weighed SVM loss function, we train our model to estimate the regression coefficient β. In this experiment, we set $\beta_s = \left\{\frac{1}{3}, -\frac{2}{3}, 1, -\frac{1}{3}, \frac{2}{3}, -1, \cdots\right\}, \beta_v = \overrightarrow{0}$, and $\varepsilon = \mathcal{N}(0, 0.3)$. In the experiment, we will set different dimension of β, hence if the dimension of β_S is higher than 6, we will set the element of which index is larger than 6 as the $i\%6$-th of β_V.

$$Y_{poly} = f(\mathbf{S}) + \varepsilon = [\mathbf{S}, \mathbf{V}] \cdot [\beta_s, \beta_v]^T + \mathbf{S}_{\cdot,1}\mathbf{S}_{\cdot,2} + \varepsilon \tag{8}$$

Baselines: We compare our model with five baseline methods:

- Ordinary Least Square (OLS) (Hutcheson, 2011):
$$\min \|Y - \mathbf{X}\beta\|_2^2$$
- Lasso (Tibshirani, 1996):
$$\min \|Y - \mathrm{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1$$
- Ridge (Hoerl & Kennard, 1970):
$$\min \|Y - \mathrm{X}\beta\|_2^2 + \lambda_1 \|\beta\|_2$$
- Decorrelated Weighting Regression (DWR) (Kuang et al., 2020b):
$$\begin{aligned}&\min_{W,\beta} \textstyle\sum_{i=1}^n W_i \cdot (Y_i - \mathbf{X}_{i,}\beta)^2 \\ \text{s.t} \quad &\textstyle\sum_{j=1}^p \left\|\mathbf{X}_{,j}^T \mathbf{\Sigma}_W \mathbf{X}_{,-j}/n - \mathbf{X}_{,j}^T W/n \cdot \mathbf{X}_{,-j}^T W/n\right\|_2^2 < \lambda_2\end{aligned}$$
- Support Vector Machines (SVM) (Suykens & Vandewalle, 1999):
$$\min_{w,b,\zeta,\zeta^*} \frac{1}{2} w^T w + \sum_{i=1}^n (\zeta_i + \zeta_i^*)$$
- SVM combined with DWR(DWR_SVM):
$$\begin{aligned}&\min_{w,b,\zeta,\zeta^*} \tfrac{1}{2} w^T w + \textstyle\sum_{i=1}^n W_i (\zeta_i + \zeta_i^*) \\ &\text{s.t } \textstyle\sum_{j=1}^p \left\|\mathbf{X}_{,j}^T \mathbf{\Sigma}_W \mathbf{X}_{,-j}/n - \mathbf{X}_{,j}^T W/n \cdot \mathbf{X}_{,-j}^T W/n\right\|_2^2 < \lambda_2\end{aligned}$$

Generating Various Environments To test the stability of the algorithms, we generate a set of environment e with a distinct distribution distribution P_{XY}. Following the Kuang's experiment Kuang et al. (2020b), we generate different environments based on various $P(S|V)$. To simplify the problem, we simulate $P(S_b|V$ on a subset $S_b \in S$, where the dimension of S_b is $0.2 * p$. We applied the bias rate equation $Pr = \prod_{\mathbf{S}_i \in \mathbf{S}_b} |r|^{-5*D_i}$ to tune the $P(S_b|V$, where $D_i = |f(\mathbf{S}) - \text{sign}(r) * \mathbf{V}_i|, r \in [-3, -1) \cup (1, 3]$. $r > 1$ indicates that Y and S_b have positive unstable relationships, while $r < -1$ corresponds to the negative unstable relationships. The higher absolute value of r the stronger connection between S_b and Y, leading to generate different environments. The result is shown in Figure 3.

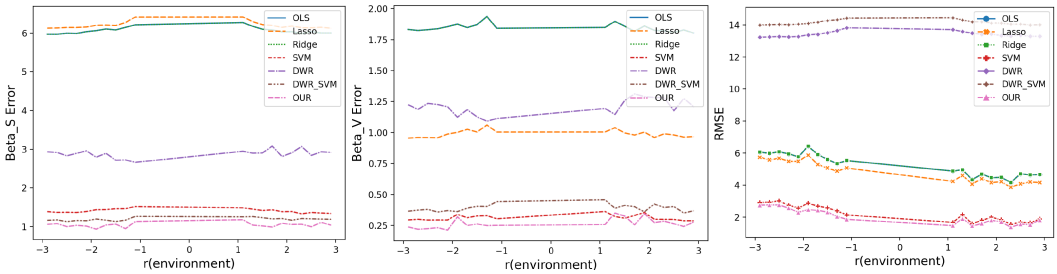

Figure 3: Figures describe the β_S, β_V and RMSE with various environments.

A.3 Causal Experiment Results and Explanation

We compare our model with five traditional methods:

- **Logistic Regression** We leverage the logistic regression classifier with L-BFGS solver for classification (Bollapragada et al., 2018).
- **Random Forest** We apply standard Random Forest classifier to solve the classification problem (Pal, 2005).
- **XGboost** We adopt XGBoost, an extreme gradient boosting methods, to compare with other models (Chen & Guestrin, 2016).
- **SVM** We apply supervised learning models, SVM, with linear kernel to analyze data for classification (Suykens & Vandewalle, 1999).
- **MLP** We use the traditional neural network multi-layer perceptron to solve this classification task (Agatonovic-Kustrin & Beresford, 2000).

We sort the rules in descending order by calculating the importance and show the top five rules compared with the doctor's score 4. The details of the description for each feature in the rules are shown in the Table 5. The scoring criteria are as follows:

- **Score 4:** Strongly agree that the rule contains causality.
- **Score 3:** Agree that the rule contains causality.
- **Score 2:** Disagree with this rule.
- **Score 1:** Strongly disagree with this rule.

Table 4: Rules filtered by algorithm are sorted in a descending order by our algorithm compared with the scores given by doctors.

Association Rules	Scores
Heart Disease	
age middle, #major vessels0, fixed defect, pressure normal, ST-T wave abnormality ⇒ heart disease	4
age middle, cholesterol edge, #major vessels0, lower than 120mg/ml ⇒ heart disease	3
non-anginal pain, cholesterol high, no exercise induced angina ⇒ heart disease	4
ST-T wave abnormality, downsloping ⇒ heart disease	4
fixed defect, #major vessels0, cholesterol edge ⇒ heart disease	4
Esophageal Cancer	
Modified Ryan Score 2.0, Esophagectomy Procedure 4 ⇒ recurrence	2
tobacco use, Alcohol Use, Neoadjuvant Radiation, Histological Grade 2, Final Histology 1 ⇒ recurrence	4
Histological Grade 3, Neoadjuvant Radiation, Esophagectomy Procedure 4, Final Histology 1 ⇒ recurrence	4
clinical m Stage 1, Histological Grade 3, Neoadjuvant Radiation, Esophagectomy Procedure 4, Final Histology 1 ⇒ recurrence	4
esoph tumor location 4, Esophagectomy Procedure 5, Histological Grade 3 ⇒ recurrence	3
Cauda Equina Syndrome	
elixsum, beds, procedure 03 09 ⇒ die360	4
Emergency, diagnosis 344 60, complication 240days ⇒ die360	4
diagnosis 344 60, life threatening, complication 240days ⇒ die360	4
if aa ⇒ die360	4
or potentially disabling conditions, complication 240days ⇒ die360	4

A.4 INTRODUCTION FOR FEATURES

Table 5: Introduction of individual features on different datasets.

Features	Explanation
Heart Disease	
age middle	Patients between the ages of 40 and 60
#major vessels0	The number of major vessels (0-3) colored by flourosopy is 0
fixed defect	Thalium stress test result is fixed defect
pressure normal	Blood pressure within the normal range
ST-T wave abnormality	Resting electrocardiography result is ST-T wave abnormality
cholesterol edge	Serum cholesterol is in range $(200, 220]$ mg/dl
lower than 120mg/ml	Fasting blood sugar is lower than 120mg/ml
non-anginal pain	Chest pain type is non-angina
cholesterol high	Serum cholesterol is higher than 220 $\mathrm{mg/dl}$
no exercise induced angina	not Exercise induced angina
downsloping	Slope of peak exercise ST segment is downsloping
heart disease	It refers to the presence of heart disease in the patient
Esophageal Cancer	
Modified Ryan Score 2.0	(near complete response): single cells or rare small groups of cancer cells
Esophagectomy Procedure 4	Complete MIS/Robotic McKeown (Three-Hole) esophagectomy
tobacco use	Use tobacco
Alcohol Use	Use Alcohol
Neoadjuvant Radiation	Patient underwent neoadjuvant radiation
Histological Grade 2	How differentiated the tumor is: Moderately Differentiated
Final Histology 1	History: Adenocarcinoma
Histological Grade 3	How differentiated the tumor is: Poorly Differentiated
clinical m Stage 1	Details any spread (metastasis) to other sites of the body: M0
esoph tumor location 4	Lower Thoracic, including GE junction
Esophagectomy Procedure 5	Hybrid (Laparoscopy + Thoracotomy) McKeown (Three-Hole) esophagectomy
recurrence	Details whether the patient experience recurrence of their cancer
Cauda Equina Syndrome	
elixsum	Elixhauser comorbidity sum for that patient is high
beds	Number of beds in the hospital is small
procedure 03 09	ICD-9-CM Procedure Codes: 03.09
Emergency	The patient requires immediate medical intervention as a result of severe
diagnosis 344 60	ICD9 indicators
complication 240days	Indicators for complication within 240 days of discharge
life threatening	The patient's condition is very dangerous
if aa	The racial of the patient is `African American`
die360	Patient died within 360 days

A.5 PROOF

Lemma 2. *If the number of features in the datasets and the terms in the Taylor expansion are fixed, when $n \to \infty$ there exists $W \succeq 0$ such that*

$$\lim_{n\to\infty} \|\mathscr{F}^{(i)}_{p_2\to p_1, i>0}\|_2^2$$

Proof. Based on our regularizer, we know that

$$\begin{pmatrix} n & \sum_i w_i x_{ip_2} & \cdots & \sum_i w_i^k x_{ip_2}^k \\ \sum_i w_i x_{ip_2} & \sum_i w_i^2 x_{ip_2}^2 & \cdots & \sum_i w_i^{k+1} x_{ip_2}^{k+1} \\ \vdots & \vdots & & \sum_i \\ \sum_i w_i^k x_{ip_2}^k & \sum_i w_i^{k+1} x_{ip_2}^{k+1} & \ddots & \sum_i w_i^{2k} x_{ip_2}^{2k} \end{pmatrix} \begin{pmatrix} f_{p_1p_2}(x_{p_2}(0)) \\ f_{p_1p_2}(x_{p_2}(0)) \\ \vdots \\ f^{(p)}_{p_1p_2}(x_{p_2}(0)) \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i w_i x_{ip_2} y_i \\ \vdots \\ \sum_i w_i^k x_{ip_2}^k y_i \end{pmatrix}$$

We assume that the covariance is 0:

$$\operatorname{cov}(\hat{x}_{ip_2}, y_i) = \operatorname{cov}\left(\hat{x}^2_{ip_2}, y_i\right) = \operatorname{cov}\left(\hat{x}^3_{ip_2}, y_i\right) = \cdots = \operatorname{cov}\left(\hat{x}^k_{ip_2}, y_i\right) = 0$$

Combine with the following equation, we can get

$$\begin{aligned} n \to \infty : \frac{1}{n}\sum_n \hat{x}_{ip_2} y_i - \frac{1}{n^2}\sum_n \hat{x}_{ip_2}\sum_n y_i &= \frac{1}{n}\sum_n \hat{x}^2_{ip_2} y_i - \frac{1}{n^2}\sum_n \hat{x}^2_{ip_2}\sum_n y_i \\ &= \frac{1}{n}\sum_n \hat{x}^k_{ip_2} y_i - \frac{1}{n^2}\sum_n \hat{x}^k_{ip_2}\sum_n y_i = 0 \end{aligned}$$

$$\begin{pmatrix} n & \sum_i \hat{x}_{ip_2} & \cdots & \sum_i \hat{x}^k_{ip_2} \\ \sum_i \hat{x}_{ip_2} & \sum_i \hat{x}^2_{ip_2} & \cdots & \sum_i \hat{x}^{k+1}_{ip_2} \\ \vdots & \vdots & & \vdots \\ \sum_i \hat{x}^k_{ip_2} & \sum_i \hat{x}^{k+1}_{ip_2} & \ddots & \sum_i \hat{x}^{2k}_{ip_2} \end{pmatrix} \begin{pmatrix} f_{p_1p_2}(x_{p_2}(0)) \\ f'_{p_1p_2}(x_{p_2}(0)) \\ \vdots \\ f^{(p)}{}_{p_1p_2}(x_{p_2}(0)) \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \sum_i y_i \hat{x}_{ip_2} \sum_i y_i \\ \vdots \\ \sum_i \hat{x}^k_{ip_2} \sum_i y_i \end{pmatrix}$$

$$\begin{pmatrix} \frac{\sum_i \hat{x}^2_{ip_2}}{\sum_i \hat{x}_{ip_2}} - \sum_i \hat{x}_{ip_2} & \frac{\sum_i \hat{x}^3_{p_2}}{\sum_i \hat{x}_{ip_2}} - \sum_i \hat{x}^2_{ip_2} & \cdots & \frac{\sum_i \hat{x}^{k+1}_{p_2}}{\sum_i \hat{x}_{ip_2}} - \sum_i \hat{x}^k_{ip_2} \\ \sum_i \hat{x}^3_{ip_2} & & & \\ \frac{\sum_i \hat{x}^2_{ip_2}}{2} - \sum_i \hat{x}_{ip_2} & \frac{\sum_i \hat{x}^4_{ip_2}}{\sum_i \hat{x}^2_{ip_2}} - \sum_i \hat{x}^2_{ip_2} & \cdots & \frac{\sum_i \hat{x}^{k+2}_{ip_2}}{\sum_i \hat{x}^2_{ip_2}} - \sum_i \hat{x}^k_{ip_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sum_i \hat{x}^{k+1}_{ip_2}}{\sum_i \hat{x}^k_{ip_2}} - \sum_i \hat{x}^k_{ip_2} & \frac{\sum_i \hat{x}^{k+2}_{ip_2}}{\sum_i \hat{x}^k_{ip_2}} - \sum_i \hat{x}^k_{ip_2} & \cdots & \frac{\sum_i \hat{x}^{2k}_{ip_2}}{\sum_i \hat{x}^k_{ip_2}} - \sum_i \hat{x}^k_{ip_2} \end{pmatrix} \begin{pmatrix} f_{p_1p_2}(x_{p_2}(0)) \\ f'' p_1p_2(x_{p_2}(0)) \\ \vdots \\ f^{(p)}_{p_1p_2}(x_{p_2}(0) \end{pmatrix} = 0$$

$$\begin{vmatrix} \frac{\sum_i \hat{x}^2_{ip_2}}{\sum_i \hat{x}_{ip_2}} - \sum_i \hat{x}_{ip_2} & \frac{\sum_i \hat{x}^3_{ip_2}}{\sum_i \hat{x}_{ip_2}} - \sum_i \hat{x}^2_{ip_2} & \cdots & \frac{\sum_i \hat{x}^{k+1}_{ip_2}}{\sum_i \hat{x}_{ip_2}} - \sum_i \hat{x}^k_{ip_2} \\ \frac{\sum_i \hat{x}^3_{ip_2}}{\sum_i \hat{x}^2_{ip_2}} - \sum_i \hat{x}_{ip_2} & \frac{\sum_i \hat{x}^4_{ip_2}}{\sum_i \hat{x}^2_{ip_2}} - \sum_i \hat{x}^2_{ip_2} & \cdots & \frac{\sum_i \hat{x}^{k+2}_{ip_2}}{\sum_i \hat{x}^2_{ip_2}} - \sum_i \hat{x}^k_{ip_2} \\ \vdots & \vdots & & \vdots \\ \frac{\sum_i \hat{x}^{k+1}_{ip_2}}{\sum_i \hat{x}^k_{ip_2}} - \sum_i \hat{x}^k_{ip_2} & \frac{\sum_i \hat{x}^{k+2}_{ip_2}}{\sum_i \hat{x}^k_{ip_2}} - \sum_i \hat{x}^k_{ip_2} & \cdots & \frac{\sum_i \hat{x}^{2k}_{ip_2}}{\sum_i \hat{x}^k_{ip_2}} - \sum_i \hat{x}^k_{ip_2} \end{vmatrix} \neq 0$$

$\hat{x}^2_{ip_2}$ is influenced by the w_i which can be adjusted, and the determinant of matrix is not equal to 0, hence the equation has only the trivial solution. We can get

$$f'_{p_1p_2}(x_{p_2}(0)) = f''_{p_1p_2}(x_{p_2}(0)) = \cdots = f^{(p)}_{p_1p_2}(x_{p_2}(0)) = 0$$

If we can prove under our regularizer, we can prove our method can work:

$$n \to \infty : (\hat{x}_{ip_2}, y_i) = \operatorname{cov}\left(\hat{x}_{ip_2}^2, y_i\right) = \operatorname{cov}\left(\hat{x}_{ip_2}^3, y_i\right) = \cdots = \operatorname{cov}\left(\hat{x}_{ip_2}^k, y_i\right) = 0$$

We set $\left(\hat{x}_{ip_2}, \hat{x}_{ip_2}^2, \ldots, \hat{x}_{ip_2}^k\right)$ is kernel density estimators: $g(x_{ip_2}$. We set the weight w_i is:

$$w_i = \frac{\prod_{iq} g\left(x_{ij}^q\right)}{\hat{G}\left(g\left(x_{i1}\right), g\left(x_{i2}\right), \ldots, g\left(x_{ip}\right)\right)}$$

$$\begin{aligned} n \to \infty : E\left[\hat{x}_{p_1}^q\right] &= \frac{1}{n} \sum_i x_{ip_1}^q \frac{\prod_{iq} g\left(x_{ij}^q\right)}{\hat{G}\left(g\left(x_{i1}\right), g\left(x_{i2}\right), \ldots, g\left(x_{ip}\right)\right)} \\ &= \int \ldots \int x_{ij}^q \prod_l g\left(x_{il}^q\right) dx_{i1} dx_{i1}^1 \ldots dx_{ip}^q + o(1) = \int x_{il}^{q_1} g\left(x_{il}^{q_1}\right) dx_{il}^{q_1} + o(1) \end{aligned}$$

$$\begin{aligned} n \to \infty : E\left[\hat{x}_{p_1}^q, \hat{x}_{p_2}\right] &= \frac{1}{n} \sum_i x_{ip_1}^q x_{ip_1} \left(\frac{\prod_{iq} g\left(x_{ij}^q\right)}{\hat{G}\left(g\left(x_{i1}\right), g\left(x_{i2}\right), \ldots, g\left(x_{ip}\right)\right)} \right)^2 \\ &= \iint x_{il}^{q_1} x_{im} g\left(x_{il}^{q_1}\right) g\left(x_{im}\right) dx_{il}^{q_1} dx_{im} + o(1) \\ &= \int x_{il}^{q_1} g\left(x_{il}^{q_1}\right) dx_{il}^{q_1} \int x_{im} g\left(x_{im}\right) dx_{im} + o(1) \end{aligned}$$

$$n \to \infty : \operatorname{cov}\left(\hat{x}_{ip_1}^q, \hat{x}_{p_1}\right) = E\left[\hat{x}_{p_1}^q\right] E\left[\hat{x}_{p_1}\right] - E\left[\hat{x}_{p_1}^q, \hat{x}_{p_2}\right] = 0$$

We can get:

$$f'_{p_1 p_2}\left(x_{p_2}(0)\right) = f''_{p_1 p_2}\left(x_{p_2}(0)\right) = \cdots = f^{(p)}_{p_1 p_2}\left(x_{p_2}(0)\right) = 0$$

□