MODEL-AGNOSTIC FEATURE SELECTION WITH ADDITIONAL MUTUAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Answering questions about data can require understanding what parts of an input $\mathbf{x}$ influence the response $\mathbf{y}$. Finding such an understanding can be built by testing relationships between variables through a machine learning model. For example, conditional randomization tests help determine whether a variable relates to the response given the rest of the variables. However, randomization tests require users to specify test statistics. We formalize a class of proper test statistics that are guaranteed to select a feature when it provides information about the response even when the rest of the features are known. We show that f-divergences provide a broad class of proper test statistics. In the class of f-divergences, the KL-divergence yields an easy-to-compute proper test statistic that relates to the additional mutual information (AMI). Questions of feature importance can be asked at the level of an individual sample. We show that estimators from the same AMI test can also be used to find important features in a particular instance. We provide an example to show that perfect predictive models are insufficient for instance-wise feature selection. We evaluate our method on several simulation experiments, on a genomic dataset, a clinical dataset for hospital readmission, and on a subset of classes in ImageNet. Our method outperforms several baselines in various simulated datasets, is able to identify biologically significant genes, can select the most important predictors of a hospital readmission event, and is able to identify distinguishing features in an image-classification task.

1 INTRODUCTION

Model interpretation techniques aim to select features important for a response by reducing models (sometimes locally) to be human interpretable. However the phrase model interpretation can be a bit of a misnomer. Any interpretation of a model must be imbued to the model by the population distribution that provides the data to train the model. In this sense, interpreting a model should be viewed as understanding the population distribution of data through the lens of a model. Existing methods for understanding the population distributions impose restrictions on the models fit to the population, choice of test statistic, or auxiliary models for interpretation (Ribeiro et al., 2016; Lundberg & Lee, 2017). Such modeling restrictions limit the applicability of these methods to a smaller class of population distributions. To be able to work in a model-agnostic manner, feature selection methods can use models but must not require a particular structure in models used to fit the data or used in interpretation.

Understanding the population distribution can be phrased as a hypothesis to be tested using samples from the population distribution in concert with a model. One hypothesis test for feature importance assesses whether a response is independent of a feature given the rest of the features; this test is called a conditional randomization test (Candes et al., 2018). Testing the importance of each feature with a conditional randomization test necessitates multiple tests. When many tests are asked of the same data, it becomes important to control the rate at which those questions might be satisfied by random chance, that is, control something like the false discovery rate (FDR) (Benjamini & Hochberg, 1995)

Hypothesis tests require test statistics. Ideally, we would like to use test statistics that make no implicit modeling assumptions. For example, using a test statistic like correlation is akin to using a linear model (Barber et al., 2015). To avoid these assumptions, we develop the notion of a *proper test statistic*. Proper test statistics reveal the true conditional independences in the population dis-

tribution. We show that expected f-divergences (Csiszár, 1964) between the joint and the marginal form a broad family of proper test statistics. In the family of f-divergences, we show that KL-divergence yields the simplest computation reducing down to regression. Using the KL-divergence in this context has a the natural interpretation; it is a measure of the additional information each feature provides about the outcome over the rest. This measure of information is known as additional mutual information (AMI) (Ranganath & Perotte, 2018). Our proposed procedure is called the additional mutual information conditional randomization test (AMI-CRT); AMI-CRT controls the FDR, works with any regression method. These regressions can be performed by using a flexible class of estimators that approximate a broad class of functions, like neural networks or random forests. Though simple, AMI-CRT outperforms popular procedures for feature importance on wide variety of simulated data, hospital records, and biological data.

Working with data sometimes requires interpreting a population distribution on the level of an individual datapoint. For example, a doctor may benefit in knowing what features for a particular patient relate to their risk of a disease. The process of identifying features at a datapoint-level is called *instance-wise* feature selection (Ribeiro et al., 2016; Lundberg & Lee, 2017; Gimenez & Zou, 2019). We identify an issue in instance-wise feature selection, where features selected with the true population distribution do not yield the features that were used to generate the response of an instance. The crux of this disparity is that response generation process conditional on the features may use randomness to select features. We provide an example to demonstrate where instance-wise feature selection can go awry. We develop sufficient conditions for instance-wise feature selection to avoid this issue, the same regression estimates from AMI-CRT can also be used to estimate feature-importances with minimal computational overhead, resulting in a method we term additional mutual information instance-wise feature selection (AMI-IW). We demonstrate AMI-IW on multiple simulations and image data. Across all of these tasks AMI-IW outperforms popular baselines.

1.1 Related Work

Permutation tests (Fisher, 1937) provide a test for marginal independence between each feature and the outcome. However, they fail in the case of conditional independence which is required when covariates are dependent on each other. To address this, solutions like Sure Independence Screening (Barut et al., 2016; Fan & Lv, 2008) and Conditional Randomization Tests (Barber et al., 2015; Candes et al., 2018) have been proposed, which give frameworks for conditional independence testing. However, these often make linearity or additive noise assumptions about the data generating distribution. Furthermore, they require the choice of a test statistic to capture some notion of conditional independence. The user of such frameworks is often burdened with the task of choosing this test statistic, which may require strong assumptions about the data generating distribution. Extending this approach to neural networks, (Lu et al., 2018) propose a fully connected network whose weights are used as a test statistic. While novel, their method is specific to fully connected networks. (Tansey et al., 2018) propose the holdout randomization test (HRT) framework which uses the empirical 0-1 risk as a test statistic, and provides computational speed-ups over CRTs.

Our contributions. *To address the issue of choosing the test statistic, we develop the notion of a proper test statistic that is necessary for model-agnostic feature selection. We show that expected f-divergences are proper test statistics and one particular choice, the* KL*-divergence, yields a computationally efficient yet effective test for conditional independence, called the* AMI-CRT.

Beyond understanding the population distribution, some tasks require interpreting a population distribution on the level of an individual datapoint. Methods that test for conditional independence work under distributional notions of feature selection, but are not designed for identifying the relevant features for a particular sample. To address this issue of "instance-wise feature selection," several methods have been proposed including local perturbations (Simonyan et al., 2013; Sundararajan et al., 2017; Ribeiro et al., 2016) and fitting simpler auxiliary models to explain the predictions of a large model (Chen et al., 2018; Lundberg & Lee, 2017; Yoon et al., 2019; Turner, 2016). Our instance-wise work is most similar to that of (Burns et al., 2019) who repurpose the HRT framework to perform instance-wise feature selection or (Gimenez & Zou, 2019) who define a conditional randomization test (CRT) procedure for subsets of the feature space. In general however, the conditions under which instance-wise feature selection with predictive models may be possible are not well developed.

Our contributions. *We address this issue by first identifying a set of sufficient conditions under which instance-wise feature selection is always possible. We then show how estimators in* AMI-CRT *can be repurposed for use in an instance-wise setting, yielding a procedure called the* AMI-IW.

2 PROPER TESTS FOR FEATURE SELECTION

Practitioners of machine learning use feature selection to identify important features for their predictive task. One way to filter out important features is to find those that improve predictions given even the rest of the features. This can be formalized through conditional independence. Let $\mathbf{x}_j$ be the j^{th} feature of $\mathbf{x}$ and let $\mathbf{x}_{-j}$ be all features but the j^{th} one. The goal is to discover a set $\mathcal{S}$ such that $\forall \mathbf{x}_j \notin \mathcal{S}, \mathbf{x}_j \perp \mathbf{y} \mid \mathbf{x}_{-j}$, where independence is with respect to the true population distribution q. The only knowledge about q comes from a finite set of samples $\mathcal{D}_N := \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^N$ sampled from the population. This means that it is impossible to assess exact conditional independence. Therefore, in the finite sample setting, we must formulate a statistical hypothesis test.

Conditional randomization tests. A conditional randomization test (CRT) (Candes et al., 2018) provides us with tools to help us define such a hypothesis test. For the jth feature, CRTs first compute some test statistic t on the N samples of data $\mathcal{D}_N$ and place this statistic in a null distribution where samples of the jth feature $\boldsymbol{x}_j^{(i)}$ have been replaced by samples of $\widetilde{\boldsymbol{x}}_j^{(i)}$ that is conditionally independent of the outcome. Letting $\widetilde{\mathcal{D}}_{j,N}$ be a dataset where $\{\boldsymbol{x}_j^{(i)}\}_{i=1}^N$ has been replaced by $\{\widetilde{\boldsymbol{x}}_j^{(i)}\}_{i=1}^N$, we can write the p-value for this test as

$$p_j(\mathcal{D}_N) = \mathop{\mathbb{E}}_{\widetilde{\boldsymbol{x}}_j^{(i)} \sim q(\mathbf{x}_j \mid \mathbf{x}_{-j} = \boldsymbol{x}_{-j}^{(i)})} \left[\mathbb{1}\left(t(\mathcal{D}_N) \leq t(\widetilde{\mathcal{D}}_{j,N}) \right) \right], \tag{1}$$

To see that $t(\widetilde{\mathcal{D}}_{j,N})$ is a valid null distribution, consider the case where $\mathbf{x}_j \perp \mathbf{y} \mid \mathbf{x}_{-j}$. Then $\widetilde{\mathbf{x}}_j \mid \mathbf{y}, \mathbf{x}_{-j}$ is equal in distribution to $\mathbf{x}_j \mid \mathbf{x}_{-j} = \mathbf{x}_j \mid \mathbf{y}, \mathbf{x}_{-j}$. Therefore, in this case, a test statistic t computed using $\mathcal{D}_N$ will have the same empirical distribution as a test statistic computed using $\widetilde{\mathcal{D}}_{j,N}$ (Candes et al., 2018). While CRTs provide a general method for conditional independence testing, they leave several components including the choice of test statistic unspecified.

2.1 CHOOSING THE RIGHT TEST STATISTIC

Imagine a test statistic $t(\cdot) = t(\{\boldsymbol{x}_j^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^N)$ that uses only a feature $\mathbf{x}_j$ and the outcome $\mathbf{y}$. Any p-values computed using this test statistic would be meaningless when testing for conditional independence, as t never considers the remaining features $\boldsymbol{x}_{-j}$. Therefore, particular choices for test statistics limit what can be tested. To address this, we introduce the concept of a *proper* test statistic.

Definition 1. ***Proper Test Statistic****: A test statistic $t(\cdot)$ is proper if p-values produced by the statistic converge to 0 when the null must be rejected, and are uniformly distributed otherwise.*

Using t in Equation (1), we should expect to see:

$$p_j\left(\mathcal{D}_N\right) \xrightarrow[N\to\infty]{d} \begin{cases} \textit{Uniform}(0,1) & \textit{if } \mathbf{x}_j \perp \mathbf{y} \mid \mathbf{x}_{-j} \\ q\left(p_j\left(\mathcal{D}_N\right) = 0\right) = 1 & \textit{if } \mathbf{x}_j \not\perp \mathbf{y} \mid \mathbf{x}_{-j} \end{cases}, \tag{2}$$

where $\xrightarrow{d}$ indicates a convergence in distribution. Under the alternate hypothesis, which in the case of feature selection is $\mathbf{x}_j \not\perp \mathbf{y} \mid \mathbf{x}_{-j}$, the power to reject the null hypothesis must be 1, implying $p_j \to 0$. A proper test statistic requires that Equation (2) must hold for all distributions of $\mathbf{y}, \mathbf{x}$.

Definition 1 mirrors the concept of a scoring rule (Gneiting & Raftery, 2007), which measures the calibration of a probabilistic prediction by a model. Generally speaking, a *proper* scoring rule is one such that the highest expected score is obtained by a model that uses the true probability distribution to make predictions. We now show that expected f-divergences may be used as a test statistic for conditional independence.

2.2 f-DIVERGENCES ARE PROPER TEST STATISTICS

f-divergences measure the closeness between two distributions and are 0 only when the two distributions are equal. Conditional independence implies that the conditional distribution factorizes:

$$q(\mathbf{x}_j, \mathbf{y} \mid \mathbf{x}_{-j}) = q(\mathbf{x}_j \mid \mathbf{x}_{-j}) q(\mathbf{y} \mid \mathbf{x}_{-j}) \tag{3}$$

This equality means that the f-divergence between the left-hand and the right-hand distributions above is 0. Using this fact, we can use f-divergences to build a proper test statistic.

Let $\hat{f}_j(\mathcal{D}_N)$ be a finite sample estimator of the true expected f-divergence $\mathbb{E}_{q(\mathbf{x}_{-j})} \mathrm{D}_f \left(q(\mathbf{x}_j, \mathbf{y} \mid \mathbf{x}_{-j}) \parallel q(\mathbf{x}_j \mid \mathbf{x}_{-j}) q(\mathbf{y} \mid \mathbf{x}_{-j}) \right)$. When $\hat{f}_j(\mathcal{D}_N)$ converges in probability as N increases to the true expected f-divergence, i.e. it is consistent, $\hat{f}_j$ is a proper test statistic.

Lemma 1. *A consistent estimator of the expected f-divergence, $\hat{f}_j(\mathcal{D}_N)$, is a proper test statistic.*

Lemma 1 implies that given sufficient data, a consistent estimator of expected f-divergence used as a test statistic in a CRT recovers the set of important features $\mathcal{S}$: those that are not conditionally independent of the response given the rest of the features. See Appendix C.2 for a proof of Lemma 1.

While a consistent estimator of an expected f-divergence is a proper test statistic, estimating f-divergences from data is a challenging task. For each feature, f-divergence in a CRT requires estimates of the following conditional distributions: $q(\mathbf{y} \mid \mathbf{x}), q(\mathbf{y} \mid \mathbf{x}_{-j})$, and $q(\mathbf{y} \mid \widetilde{\mathbf{x}}_j, \mathbf{x}_{-j})$. To avoid direct estimation of these conditionals, (Nowozin et al., 2016) use adversarial techniques to estimate f-divergences. However, adversarial techniques suffer from stability issues during training (Arjovsky et al., 2017; Salimans et al., 2016).

In the next section, we discuss how a specific choice of the f-divergence, the KL-divergence, can help simplify this estimation task.

2.3 ADDITIONAL MUTUAL INFORMATION CONDITIONAL RANDOMIZATION TEST (AMI-CRT)

Recall that $\mathcal{D}_N := \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^N$ is a finite sample from some population distribution q. Let $\widetilde{\mathcal{D}}_{j,N} := \{(\widetilde{\boldsymbol{x}}_j^{(i)}, \boldsymbol{x}_{-j}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^N$ be a dataset where the jth feature of $\mathcal{D}_N$ is replaced with a feature that is conditionally independent of the outcome. To compute the p-value in a CRT we use an empirical estimate $\hat{f}_j(\mathcal{D}_N)$ of the expected f-divergence. For some convex function f, let $f^* = tf(\frac{1}{t})$ and compute the f^*-divergence (Sason & Verdu, 2016) which can be evaluated using samples from the joint data distribution. Then we need to evaluate the following difference:

$$\hat{f}_j(\mathcal{D}_N) - \hat{f}_j(\widetilde{\mathcal{D}}_{j,N}) = \frac{1}{N} \sum_{i=1}^N f \left(\frac{q(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)})}{q(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}_{-j}^{(i)})} \right) - \frac{1}{N} \sum_{i=1}^N f \left(\frac{q(\boldsymbol{y}^{(i)} \mid \widetilde{\boldsymbol{x}}_j^{(i)}, \boldsymbol{x}_{-j}^{(i)})}{q(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}_{-j}^{(i)})} \right). \tag{4}$$

For each feature $\mathbf{x}_j$, we would therefore need estimates for three distributions: $q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})$, $q(\mathbf{y} \mid \widetilde{\mathbf{x}}_j, \mathbf{x}_{-j})$, and $q(\mathbf{y} \mid \mathbf{x}_{-j})$. For a particular choice of f, we can actually reduce the number of distributions to be estimated. If we let $f(z) = z \log z$, i.e. the KL-divergence, we can avoid having to estimate $q(\mathbf{y} \mid \mathbf{x}_{-j})$, simplifying Equation (4) into:

$$\frac{1}{N} \sum_{i=1}^N \log q(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}) - \frac{1}{N} \sum_{i=1}^N \log q(\boldsymbol{y}^{(i)} \mid \widetilde{\boldsymbol{x}}_j^{(i)}, \boldsymbol{x}_{-j}^{(i)}). \tag{5}$$

The KL-divergence between $q(\mathbf{y}, \mathbf{x}_j \mid \mathbf{x}_{-j})$ and $q(\mathbf{y} \mid \mathbf{x}_{-j}) q(\mathbf{x}_j \mid \mathbf{x}_{-j})$, called the additional mutual information (AMI) (Ranganath & Perotte, 2018), helps cancel out the $q(\mathbf{y} \mid \mathbf{x}_{-j})$ term when used in a CRT. We show the full derivation of this in Appendix D.

Avoiding the estimation of $q(\mathbf{y} \mid \mathbf{x}_{-j})$ provides a significant computational advantage. This helps control the sources of error in our p-value computation. By Lemma 1, if we have a consistent estimator for the quantities in Equation (5), we can be sure that this simplified test statistic is proper.

Estimating *p*-values with flexible regression models. As discussed earlier, importance is an attribute of the population distribution. If we had access to the true conditional distribution $q(\mathbf{y} \mid \mathbf{x})$, we might be able to determine feature importance by directly using Equation (5). However, this is

Algorithm 1: AMI-CRT for feature selection

Input: $\boldsymbol{x} \in \mathbb{R}^{N \times D}$, feature matrix; $\boldsymbol{y} \in \mathbb{R}^N$, labels; j, the jth feature to be tested
Output: p_j, the p-value for the null hypothesis of $\mathbf{x}_j \perp \mathbf{y} \mid \mathbf{x}_{-j}$
Fit regression $q_\theta(\boldsymbol{x}_j \mid \boldsymbol{x}_{-j})$
Fit regression $q_\beta(\boldsymbol{y} \mid \boldsymbol{x})$
Let $t_j = \frac{1}{N} \sum_{i=1}^N \log q_\beta(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x} = \boldsymbol{x}^{(i)})$
for $k \in [1, 2, \ldots, K]$ **do**
  Let $\{\widetilde{\boldsymbol{x}}_j^{(i)}\}_{i=1}^N$ be a dataset such that $\widetilde{\boldsymbol{x}}_j^{(i)} \sim q_\theta(\mathbf{x}_j \mid \mathbf{x}_{-j} = \boldsymbol{x}_{-j})$
  Let $\widetilde{\boldsymbol{x}}^{(k)} := \{\widetilde{\boldsymbol{x}}_j^{(i)}, \boldsymbol{x}_{-j}^{(i)}\}_{i=1}^N$
  Fit regression $q_\beta^{(k)}\left(\widetilde{\boldsymbol{x}}^{(k)}\right) = \boldsymbol{y}$
  Let $\tilde{t}_j^{(k)} = \frac{1}{N} \sum_{i=1}^N \log q_\beta^{(k)}(\boldsymbol{y}^{(i)} \mid \widetilde{\boldsymbol{x}}_j^{(i)}, \boldsymbol{x}_{-j}^{(i)})$
end
Let $p_j = \frac{1}{K+1}\left[1 + \sum_{k=1}^K \mathbb{1}\left(t_j \leq \tilde{t}_j^{(k)}\right)\right]$

usually not the case and we must estimate them from data. This is where models of the world can be useful. Using an appropriate model, we can efficiently estimate the true population distributions q. While using a restrictive model class can potentially prevent a test statistic from being proper, modeling assumptions may be the easiest way to consistently estimate the conditional distributions we care about from data. For example, in cases where the number of features is far greater than the number of samples, a sparse model can potentially provide a better estimate for $q(\mathbf{y} \mid \mathbf{x})$ than a non-parametric model. Since we use AMI, we can use models to compute two regressions: one from $\mathbf{x}$ to $\mathbf{y}$, and one from $\{\widetilde{\mathbf{x}}_j, \mathbf{x}_{-j}\}$ to $\mathbf{y}$. These models serve as estimators for both $q(\mathbf{y} \mid \mathbf{x})$ and $q(\mathbf{y} \mid \widetilde{\mathbf{x}}_j, \mathbf{x}_{-j})$.

We also use models to help generate samples from $\widetilde{\mathbf{x}}_j$, a random variable designed to be independent of $\mathbf{y}$ given $\mathbf{x}_{-j}$. As $\widetilde{\mathbf{x}}_j \sim q(\mathbf{x}_j \mid \mathbf{x}_{-j})$, we can again estimate this distribution using regression from $\mathbf{x}_{-j}$ to $\mathbf{x}_j$. We discretize the $\mathbf{x}_j$ being estimated, and learn parameters $\theta = \{\theta_1, \ldots, \theta_T\}$ for:

$$\boldsymbol{x}_j^{(i)} \sim \text{Categorical}(g_0(\boldsymbol{x}_{-j}^{(i)}, \theta_1), \ldots, g_T(\boldsymbol{x}_{-j}^{(i)}, \theta_T)) \tag{6}$$

where T is the number of bins. We refer the reader to Proposition 2 of (Miscouridou et al., 2018) for proof that this approximation converges in distribution to the true $q(\mathbf{x}_j|\mathbf{x}_{-j})$ as the bin width $w \to 0$ and $Nw \to \infty$.

We give an AMI-CRT procedure in Algorithm 1. We first fit a model for $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$, and a model for $q(\mathbf{y} \mid \mathbf{x})$. We then compute the expected log-probability of $\mathbf{y} \mid \mathbf{x}$. K datasets are sampled from our estimate of $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$, and used to estimate the null distribution for the expected log-probability. This is used to then compute a p-value for the conditional independence test of $\mathbf{x}_j \perp \mathbf{y} \mid \mathbf{x}_{-j}$. We also present a k-fold version of our method for higher power in Appendix B. This algorithm splits dataset $(\boldsymbol{x}, \boldsymbol{y})$ into several folds, then uses all but one of the folds to estimate the relevant conditional distributions. The remaining fold is used to estimate the expected log-probability difference. Finally, to combine the folds, we average the estimated log-probability from each fold before computing p-values. To estimate each of q_θ and q_β, standard regression models like logistic regression, neural networks, and random forests can be used at no more computational cost than training. The estimation procedure is straightforward, but effective as we demonstrate in Section 4.

An additional benefit of AMI-CRT is the ability to reuse this empirical framework to provide feature importances on an instance-wise level.

3 Tests for Instance-Wise Feature Selection

So far, we recover features that are important across the whole population, i.e. the problem of distributional feature selection. We develop a test for distributional feature selection, but do not yet address the issue that different samples could have different important features. We call this problem of recovering important features for each sample, *instance-wise feature selection (*IWFS*)*.

To identify important features instance-wise, we can use the probability of observing a particular label $\boldsymbol{y}^{(i)}$ given a set of features $\boldsymbol{x}^{(i)}$. This suggests a candidate definition for important features in IWFS:

Definition 2. ***(Candidate) Feature importance in*** **IWFS***: Let $\mathcal{D}$ be a data-generating distribution. Let $\left\{\left(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\right)\right\}_{i=1}^{N}$ be a dataset where each $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \sim \mathcal{D}$. The jth feature for the ith sample, $\boldsymbol{x}_j^{(i)}$, is an important feature if*

$$q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x} = \boldsymbol{x}^{(i)}) > q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x}_{-j} = \boldsymbol{x}_{-j}^{(i)})$$

where q is the true probability distribution specified by $\mathcal{D}$.

Definition 3 says that a feature $\boldsymbol{x}_j^{(i)}$ is important if observing it increases the probability of $\boldsymbol{y}^{(i)}$. This formulation is exploited in (Yoon et al., 2019) to obtain instance-wise important features. However, Definition 3 can sometimes fail to identify relevant features, even with access to the true conditional distribution $q(\mathbf{y} \mid \mathbf{x})$ as specified by the data generating process. While important features may satisfy this condition, so will a few unimportant features. As a demonstrative example, consider the data generating process where $\mathbf{y} = \mathbf{z}\mathbf{x}_1 + (1 - \mathbf{z})\mathbf{x}_2 + \epsilon$, $\mathbf{z} \sim \text{Bernoulli}(0.5)$, $\mathbf{x}_1, \mathbf{x}_2 \sim \mathcal{N}(0, \sigma_{\mathbf{x}}^2)$, and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Assume we have the true $q(\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_2)$, and let $\mathbf{z}$ be unobserved. Pick a sample $(\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \boldsymbol{y}^{(i)})$ where the corresponding $\boldsymbol{z}^{(i)} = 1$, meaning that $\boldsymbol{x}_1^{(i)}$ is important for this example. We can expand the difference $p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}) - p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)})$ as:

$$p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}) - p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}) = \frac{1}{2}\mathcal{N}\left(\boldsymbol{y}^{(i)}; \boldsymbol{x}_2^{(i)}, \sigma_\epsilon^2\right) - \frac{1}{2}\mathcal{N}\left(\boldsymbol{y}^{(i)}; 0, \sigma_{\boldsymbol{x}}^2 + \sigma_\epsilon^2\right)$$

For all i such that $\boldsymbol{y}^{(i)}$ lies in a non-0 interval around $\boldsymbol{x}_2^{(i)}$, we have that $p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}) - p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}) > 0$. For example let $\sigma_\epsilon = \sigma_x = 1$, then $\boldsymbol{x}_2^{(i)} = 5$, we have that $\boldsymbol{y} \in [3, 7]$ satisfies this. In all of those cases, the wrong feature will be selected as important as per the candidate Definition 3. We show the full derivation of this example in Appendix E.1. The fundamental issue with the formulation in Definition 3 is that noise can act as a "selection" mechanism, but cannot be estimated because it is unobserved. Therefore, while it suffices to build predictive models q_θ of $q(\mathbf{y} \mid \mathbf{x})$ for distributional feature selection, IWFS might not be possible even with access to $q(\mathbf{y} \mid \mathbf{x})$.

3.1 Properties of instance-wise feature selection tests

We develop the following sufficient condition under which to construct a test for instance-wise feature importance:

Definition 3. ***Sufficient conditions for instance-wise feature selection****: Let $\mathcal{S}^{(i)}$ be a set of features for each sample $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$ such that:*

$$\mathcal{S}^{(i)} = \left\{\boldsymbol{x}_j^{(i)} : q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x} = \boldsymbol{x}^{(i)}) > q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x}_{-j} = \boldsymbol{x}_{-j}^{(i)})\right\}. \tag{7}$$

If $\boldsymbol{y}$ is discrete, and $p(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x} = \boldsymbol{x}^{(i)}) = 1$ for each sample $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$, i.e. we have perfect predictions on our dataset, then it is possible to determine which features contribute to the prediction of $\boldsymbol{y}$, instance-wise.

The set in Equation (7) consists of only features $\boldsymbol{x}_j^{(i)}$ that help increase the likelihood of observing $\boldsymbol{y}^{(i)}$ given the remaining features $\boldsymbol{x}_{-j}^{(i)}$. If the perfect predictions property of $q(\mathbf{y} \mid \mathbf{x})$ is true, then $q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x}_{-j} = \boldsymbol{x}_{-j}^{(i)})$ can only be less than or equal to $q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x} = \boldsymbol{x}^{(i)})$, with equality when $\boldsymbol{x}_j^{(i)}$ is not important to $\boldsymbol{y}^{(i)}$. Assuming the sufficient conditions in Definition 3, we can now construct an instance-wise feature selection procedure using the same estimators from AMI-CRT.

3.2 Additional mutual information instance-wise feature selection (AMI-IW)

Instance-wise feature selection can be performed using complete conditional knockoffs (CCKs). Starting from Definition 2, we begin by manipulating $q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \boldsymbol{x} = \boldsymbol{x}^{(i)})$ and marginalizing out

Feature	ami-crt	loss-hrt	corr-crt	lime	shap	rf
p-values	✓	✓	✓			
Well calibrated p-values	✓					
FDR control	✓	✓	✓			
Instance-wise feature selection	✓	✓		✓	✓	
No distributional assumptions	✓	✓		✓	✓	✓

Table 1: AMI-CRT produces well-calibrated p-values, provides FDR control, allows instance-wise feature selection, and makes no distributional assumptions about the data-generating process. This table compares popular feature selection methods to AMI-CRT.

$\boldsymbol{x}_j^{(i)}$. We then use Jensen's inequality to upper-bound the log of this expectation as follows:

$$\mathop{\mathbb{E}}_{\widetilde{\boldsymbol{x}}_j^{(i)} \sim q(\mathbf{x}_j \mid \boldsymbol{x}_{-j}^{(i)})} \left[-\log q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x}_j = \widetilde{\boldsymbol{x}}_j, \mathbf{x}_{-j} = \boldsymbol{x}_{-j}^{(i)}) \right] \geq -\log q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x}_{-j} = \boldsymbol{x}_{-j}^{(i)}) \quad (8)$$

This suggests the following instance-wise test. If the inequality in Equation (8) is strict, the feature is considered important. If equality holds in Equation (8), the feature is considered unimportant. Notice that Jensen's inequality could introduce slack in this bound that could make a feature seem relevant when it is not. We use Definition 3 to show that this is not an issue.

Recall that given a model q_θ for $q(\mathbf{y} \mid \mathbf{x})$ which satisfies the instance-wise sufficient conditions in Definition 3, $q_\theta(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x} = \boldsymbol{x}^{(i)}) \geq q_\theta(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x}_{-j} = \boldsymbol{x}_{-j}^{(i)})$. In the case where $\boldsymbol{x}_j^{(i)}$ does not help predict $\boldsymbol{y}^{(i)}$, $q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x}_j = \widetilde{\boldsymbol{x}}_j^{(i)}, \mathbf{x}_{-j} = \boldsymbol{x}_{-j}^{(i)})$ cannot be greater than $q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x}_{-j} = \boldsymbol{x}_{-j}^{(i)})$, as the former does not depend on $\widetilde{\boldsymbol{x}}_j^{(i)}$. Then the left-hand side of Equation (8) becomes $-\log q(\mathbf{y} = \boldsymbol{y}^{(i)} \mid \mathbf{x}_{-j} = \boldsymbol{x}_{-j}^{(i)})$, a constant with respect to $\widetilde{\boldsymbol{x}}^{(i)}$, implying a strict equality. Therefore, checking for equality in Equation (8) is a valid test to see if a feature is either important or unimportant. In Appendix E.2, we detail an example which shows how scores computed using Equation (8) can help rank features from most to least helpful for prediction. We term this scoring procedure for each feature the additional mutual information instance-wise feature selection (AMI-IW).

If we computed an expectation over $\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}$ of Equation (8), this procedure resembles AMI-CRT. This means we can reuse the estimators from our distributional feature selection procedure to compute these instance-wise log-probability differences for AMI-IW.

4 Experiments

We compare our method, the AMI-CRT [ami-crt] to widely-used approaches on the basis of FDR (where applicable), the area under a receiver operating characteristic (ROC) curve, and on precision and recall in the case of instance-wise feature selection:

- Correlation [corr-crt]: Difference between Correlation($\mathbf{x}_j, \mathbf{y}$) and Correlation($\widetilde{\mathbf{x}}_j, \mathbf{y}$) as a test statistic for a CRT.
- Zero-one binary classification loss [loss-hrt]: Binary classification loss difference between $\hat{\pi}_A(\boldsymbol{y} \mid \boldsymbol{x})$ and $\hat{\pi}_B(\boldsymbol{y} \mid \widetilde{\boldsymbol{x}}_j, \boldsymbol{x}_{-j})$ where $\hat{\pi}_A$ and $\hat{\pi}_B$ are models for $\boldsymbol{y} \mid \boldsymbol{x}$ as a test statistic used in a HRT (Tansey et al., 2018)
- Local interpretable model-agnostic explanations (LIME) [lime] (Ribeiro et al., 2016)
- Shapley additive explanations (SHAP) [shap] (Lundberg & Lee, 2017)
- Random forest [rf] feature importances scores

We present a summary comparison of these methods in Table 1.

4.1 Simulated data experiments

We simulate data for testing the performance of various test statistics in both a distributional feature selection setting, and in an instance-wise feature selection setting.

Dataset	ami-crt	loss-hrt	corr-crt	lime	shap	rf
orange	**0.97**	0.94	0.22	0.94	0.95	0.94
xor	**1.00**	0.95	0.45	**1.00**	0.99	0.95

Table 2: Simulated data results: Here we use the scores provided by each method to select features. We observe that the area under the receiver operating characteristic (ROC) curve for AMI-CRT is greater than that of state-of-the-art methods.

XOR [xor]: To test the case where features on their own are not informative, but together provide information, we use the xor dataset. We first sample $\mathbf{x} \sim \mathcal{N}(0, \Sigma_D)$ N times, where Σ_D is a D-dimensional covariance matrix. We randomly translate the first two dimensions of each sample $\boldsymbol{x}^{(i)}$ away from the origin in 4 different directions : $\{(s, s), (-s, s), (s, -s), (-s, -s)\}$ with labels $\{1, 0, 0, 1\}$ respectively, where $s \in \mathbb{R}$ is some fixed constant. The remaining features are conditionally independent of $\mathbf{y}$.

Orange skin (Chen et al., 2018) [orange]: To test the case where $\mathbf{y}$ is some nonlinear function of $\mathbf{x}$, we use the orange dataset. In this dataset, $\mathbf{x} \sim \mathcal{N}(0, \Sigma_D)$, $\mathbf{y} \mid \mathbf{x} \sim \exp\left(\sum_{j=1}^{\ell} \mathbf{x}_j^2 - \ell\right)$, where $\ell < D$ is the number of important features.

Selector [selector]: This experiment tests instance-wise feature selection methods. We first sample $\mathbf{x} \sim \mathcal{N}(0, \Sigma_D)$ N times, where Σ_D is a D-dimensional covariance matrix, and $D \geq 11$. The first feature $\mathbf{x}_1$, called the "selector" feature, determines the feature selection mechanism. With σ denoting the sigmoid function, we generate $\mathbf{y} \in \{0, 1\}$ as:

$$q(\mathbf{y} = 1 \mid \mathbf{x}) = \begin{cases} \sigma(\langle \beta_1, \mathbf{x}_{2:6} \rangle) & \text{if } \mathbf{x}_1 > 0 \\ \sigma(\langle \beta_2, \mathbf{x}_{7:11} \rangle) & \text{if } \mathbf{x}_1 \leq 0 \end{cases}$$

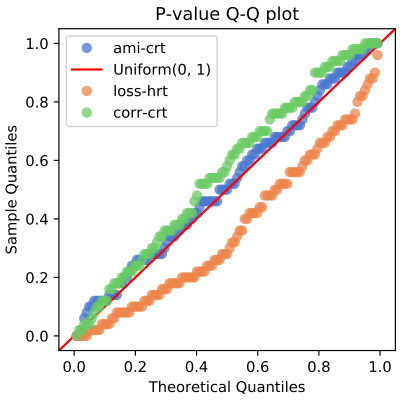

Figure 1: Quantile-Quantile plot showing uniformity of p-values for each FDR-controlling method.

Results. For *distributional feature selection* algorithms based on CRTs or HRTs, we select features using p-values. For the baselines that do not produce p-values, we select features using importance scores as determined by each method. By thresholding either p-values or importance scores, we compute an ROC curve for each method. We present the mean area under each curve over 100 simulations for the xor and orange datasets in Table 2. We notice that the while the baselines perform quite well, ami-crt achieves a higher area under the ROC.

We also investigate the calibration of p-values between ami-crt, corr-crt, and loss-hrt. We don't include other baselines in this comparison as they do not produce p-values. As stated in Definition 1, a proper test statistic must yield p-values that are uniformly distributed in the null case, and tend to 0 otherwise. We notice that in the case where a feature is important, both ami-crt and loss-hrt yield p-values close to 0, while corr-crt does not, as reflected by Table 2. Figure 1 shows a quantile-quantile plot of null p-values. We notice that ami-crt produces null p-values that are very well calibrated, as the empirical quantiles are almost the same as the expected quantiles for a uniform distribution. loss-hrt performs well in terms of area under the ROC curve, but produces p-values that are generally deflated, making it more likely for features to be incorrectly identified as important. We include additional comparison to using AMI as a test statistic in the HRT framework (Tansey et al., 2018). HRTs provide further computational speedup by using the same model for $q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})$ and $q(\mathbf{y} \mid \widetilde{\mathbf{x}}_j, \mathbf{x}_{-j})$ and thus can be evaluated faster than CRTs. However, refitting the estimators which use $\widetilde{\mathbf{x}}_j$ can yield more conservative p-values and less variance in a randomiza-

tion test. We explore this tradeoff further and discuss results of AMI as a test statistic in an HRT in Appendix F.

In the case of *instance-wise feature selection*, we evaluated methods in the following manner. For each sample in a held-out portion of the `selector` data, we looked at the percent of important features identified by each method. Table 4 (Appendix G) shows the average of these percentages across each sample of the held-out set. We notice that `ami-crt` achieves the highest recall in identifying important features compared to the baselines. Further, we observe that while `lime` and `shap` rarely identify the selector variable, `ami-crt` identifies the selector variable in nearly all of the samples as shown in Table 5 (Appendix G).

4.2 GENOMICS EXPERIMENTS

Wellcome Trust Celiac disease: To evaluate our framework on a real-world example, we use data from a genomic analysis on Celiac disease (Dubois et al., 2010). For each individual in this dataset, we have a set of single nucleotide polymorphisms (SNPs). SNPs represents genetic variance in the individual with respect to some reference population. This dataset consists of two classes of individuals: cases ($n = 3796$) and controls ($n = 8154$), where the cases are those with Celiac disease. After standard preprocessing steps as prescribed by Bush & Moore (2012), we end up with 1759 SNPs. To model $q(\mathbf{x}_j|\mathbf{x}_{-j})$, we use the same procedures as (Candes et al., 2018) where we estimate the distribution $q(\mathbf{x}_j|\mathbf{x}_{\gamma_j})$ where γ_j is only the set of SNPs (not including $\mathbf{x}_j$) known to be correlated with $\mathbf{x}_j$. To model $q(\mathbf{y}|\mathbf{x})$, we use a logistic regression model with an L_1 penalty.

Results. In Table 6 (Appendix H), we show the results of each FDR-controlling feature selection method to identify the SNPs that most likely contribute to distinguishing between those with Celiac disease and those without it. Since these method produce p-values, we control the FDR at 20% using the Benjamini & Hochberg (1995) procedure. We report the percentage of selected SNPs that have been previously shown to be associated with Celiac disease in a biological context as reported by one of (Dubois et al., 2010; Sollid, 2002; Adamovic et al., 2008; Hunt et al., 2008). We notice that `ami-crt` outperforms all other methods tested. We also list the SNPs returned by `ami-crt` in Appendix H. Since this is a real-world genomic dataset, there is no ground truth information, and as such we cannot compute ROC curves for methods that do not produce p-values.

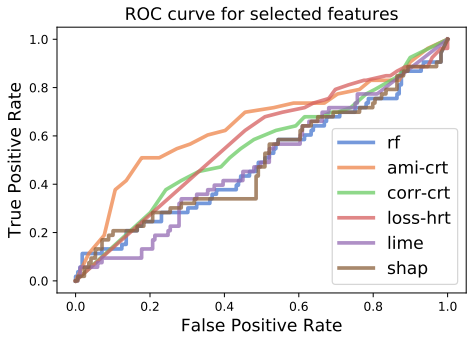

Figure 2: Using results from (Strack et al., 2014) as a clinically validated ground truth, we observe that AMI-CRT is able to achieve the highest area under the ROC curve when compared to state-of-the-art benchmarks.

4.3 HOSPITAL READMISSION EXPERIMENTS

Hospital readmission: We use a dataset consisting of ten years of medical logs from over 130 hospitals. Features in the dataset include time spent in the hospital, medical specialty of attending doctor, age, and various other diagnostic information. Labels for each sample represent one of three events: readmitted within the next 30 days ($n = 35,545$), readmitted after 30 days ($n = 11,357$), or not readmitted ($n = 54,864$). Due to class imbalance, we grouped all readmitted patients into one category ($n = 46,902$). To model $p(\boldsymbol{y}|\boldsymbol{x})$, we use a random forest classifier with 100 estimators. To model $p(\boldsymbol{x}_j|\boldsymbol{x}_{-j})$, we used the regression approach highlighted in Equation (6).

Results We first identified ground truth features as mentioned in (Strack et al., 2014). We then use the scores for each method and our ground truth features to compute an ROC curve. Figure 2 shows these curves for each method. We observe that `ami-crt` achieves a higher area under the ROC curve than state-of-the-art approaches. We discuss preprocessing steps and further results in more detail in Appendix I.

Figure 3: Instance-wise feature selection using AMI. The first and third columns show the original image of ambulances or policevans respectively. The second and fourth columns show only the patches which were found to have non-zero AMI with the label, given the rest of the patches.

4.4 IMAGENET EXPERIMENTS

To evaluate AMI-IW in the context of instance-wise feature selection, we use a subset of ImageNet. We consider the task of differentiating between ambulances and policevans. This task is interesting as both objects are physically very similar and there are only a few indicators that can be used to differentiate the two. Rather than consider each pixel as an individual feature $\boldsymbol{x}_j$, we consider a patch of pixels $\boldsymbol{x}_S$ as a single feature, such that no two patches contain overlapping pixels. To model the distribution $p(\boldsymbol{x}_S|\boldsymbol{x}_{-S})$, we make use of a generative inpainting model π_g (Yu et al., 2018). To model $p(\boldsymbol{y}|\boldsymbol{x})$, we use the VGG-16 network (Simonyan & Zisserman, 2015). To perform our instance-wise test, we compute log-probability differences using fifty generated samples from $p(\boldsymbol{x}_S|\boldsymbol{x}_{-S})$.

Results In Figure 3, we show a subset of results of instance-wise AMI-CRT. The first and third columns show the original images for each class: ambulance and policevan respectively. The second and fourth columns mask out the original image in patches where the patch is not found to be relevant to prediction. The model used to estimate $p(\boldsymbol{y}|\boldsymbol{x})$ is able to achieve roughly 90% accuracy on a held-out test set. We see that our predictive model uses relevant details like the words "ambulance" or "police" printed on the vehicle to distinguish between each class. The model also tends to ignore objects like windscreens and other features shared across classes. These results indicate that the difference in log probabilities between a model using the true data, and one using CCKs works well in determining a relevant set of features even on an instance-wise level. We show several additional images in Figure 5, in Appendix J. We also compare our method to local interpretable model-agnostic explanations (LIME) and shapley additive explanations (SHAP) (Figures 6 and 7). Both methods perform reasonably well on this task, but identify objects that are known to be common to both classes like wheels and headlamps. Neither method identifies writing on the vehicles in the images.

5 DISCUSSION

In this paper, we make the important distinction between interpreting models and understanding the population distribution of data through the lens of a model. We propose doing the latter by testing for conditional independence of each feature $\mathbf{x}_j \perp \mathbf{y} \mid \mathbf{x}_{-j}$ from a finite sample from the population distribution. However, doing so requires the careful selection of a test statistic, which is unspecified in popular feature selection frameworks. We propose a new class of model-agnostic test statistics called *proper* test statistics that guarantee control of the false discovery rate (FDR). We show that expected f-divergences between the joint distribution $q(\mathbf{x}_j, \mathbf{y} \mid \mathbf{x}_{-j})$ and the product of marginals $q(\mathbf{x}_j \mid \mathbf{x}_{-j})q(\mathbf{y} \mid \mathbf{x}_{-j})$ form a broad class of proper test statistics. By choosing the KL-divergence, and computing the AMI, we can reduce the computation of our method, the AMI-CRT, to a series of regressions. We also extend this framework to develop the AMI-IW, an instance-wise feature selection method that can help interpret a population distribution on the level of a single datapoint. We observe that both the AMI-CRT and AMI-IW outperform several popular methods in various simulated tasks, in identifying biologically significant genes, selecting the most indicative features to predict hospital readmissions, and in identifying distinguishing features in an image classification task.

REFERENCES

Svetlana Adamovic, SS Amundsen, BA Lie, AH Gudjonsdottir, H Ascher, J Ek, DA Van Heel, S Nilsson, LM Sollid, and Å Torinsson Naluai. Association study of il2/il21 and fcgriia: significant association with the il2/il21 region in scandinavian coeliac disease families. *Genes and immunity*, 9(4):364, 2008.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223, 2017.

Rina Foygel Barber, Emmanuel J Candès, et al. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Emre Barut, Jianqing Fan, and Anneleen Verhasselt. Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277, 2016.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Collin Burns, Jesse Thomason, and Wesley Tansey. Interpreting black box models with statistical guarantees. *arXiv preprint arXiv:1904.00045*, 2019.

William S Bush and Jason H Moore. Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. Learning to explain: An information-theoretic perspective on model interpretation. *arXiv preprint arXiv:1802.07814*, 2018.

Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1964.

Patrick CA Dubois, Gosia Trynka, Lude Franke, Karen A Hunt, Jihane Romanos, Alessandra Curtotti, Alexandra Zhernakova, Graham AR Heap, Róza Ádány, Arpo Aromaa, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4): 295, 2010.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Ronald Aylmer Fisher. *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937.

Jaime Roquero Gimenez and James Zou. Discovering conditionally salient features with statistical guarantees. *arXiv preprint arXiv:1905.12177*, 2019.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Karen A Hunt, Alexandra Zhernakova, Graham Turner, Graham AR Heap, Lude Franke, Marcel Bruinenberg, Jihane Romanos, Lotte C Dinesen, Anthony W Ryan, Davinder Panesar, et al. Novel celiac disease genetic determinants related to the immune response. *Nature genetics*, 40(4):395, 2008.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Yang Lu, Yingying Fan, Jinchi Lv, and William Stafford Noble. Deeppink: reproducible feature selection in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 8676–8686, 2018.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

Xenia Miscouridou, Adler Perotte, Noemie Elhadad, and Rajesh Ranganath. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, pp. 244–256, 2018.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pp. 271–279, 2016.

Rajesh Ranganath and Adler Perotte. Multiple causal inference with latent confounding. *arXiv preprint arXiv:1805.08273*, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pp. 1135–1144. ACM, 2016.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pp. 2234–2242, 2016.

Igal Sason and Sergio Verdu. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Ludvig M Sollid. Coeliac disease: dissecting a complex inflammatory disorder. *Nature Reviews Immunology*, 2(9):647, 2002.

Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328. JMLR. org, 2017.

Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M Blei. The holdout randomization test: Principled and easy black box feature selection. *arXiv preprint arXiv:1811.00645*, 2018.

Ryan Turner. A model explanation system. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2016.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. INVASE: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJg_roAcK7.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5505–5514, 2018.

A CRT ALGORITHM

We show the general CRT procedure first introduced by (Candes et al., 2018) in Algorithm 2.

Algorithm 2: Conditional Randomization Test

Input: $\boldsymbol{x} \in \mathbb{R}^{N \times D}$, feature matrix; $\boldsymbol{y} \in \mathbb{R}^N$, labels
Output: $\mathbf{p}$, the p-values for $\boldsymbol{x}_j \perp \boldsymbol{y} \,|\, \boldsymbol{x}_{-j}\ \forall j$
for $j \in [1, 2, \ldots, D]$ **do**
  Let $t_j = t(\boldsymbol{x}_j, \boldsymbol{x}_{-j}, \boldsymbol{y})$
  for $k \in [1, 2, \ldots, K]$ **do**
    Let $\widetilde{\boldsymbol{x}}^{(k)}$ be a dataset such that $\widetilde{\boldsymbol{x}}_{-j} = \boldsymbol{x}_{-j}$, and $\widetilde{\boldsymbol{x}}_j$ is randomly sampled from $p(\boldsymbol{x}_j \,|\, \boldsymbol{x}_{-j})$
    Let $\tilde{t}_j^{(k)} = t(\widetilde{\boldsymbol{x}}_j^{(k)}, \boldsymbol{x}_{-j}, \boldsymbol{y})$
  end
  Let $\mathbf{p}_j = \frac{1}{K+1}\left[1 + \sum_{k=1}^{K} \mathbb{1}\left(t_j \leq \tilde{t}_j^{(k)}\right)\right]$
end

B AMI-CRT ALGORITHM

We show the AMI-CRT procedure in Algorithm 3.

C PROOFS AND DERIVATIONS

C.1 AMI-CRT PRODUCES VALID p-VALUES

Lemma 2. *If we construct a* CCK $\tilde{\mathbf{x}}_j$ *such that* $p(\tilde{\mathbf{x}}_j \,|\, \mathbf{y}, \mathbf{x}_{-j}) = p(\mathbf{x}_j \,|\, \mathbf{x}_{-j})$*, then under the null hypothesis* $H_0 := \mathbf{x}_j \perp \mathbf{y} \,|\, \mathbf{x}_{-j}$*,* $\hat{I}(\{\mathbf{x}_j^{(i)}, \mathbf{y}^{(i)}, \mathbf{x}_{-j}^{(i)}\}_{i=1}^N)$ *is equal in distribution to* $\hat{I}(\{\widetilde{\mathbf{x}}_j^{(i)}, \mathbf{y}^{(i)}, \mathbf{x}_{-j}^{(i)}\}_{i=1}^N)$.

Proof. Under H_0, $p(\mathbf{x}_j \,|\, \mathbf{x}_{-j}, \mathbf{y}) = p(\mathbf{x}_j \,|\, \mathbf{x}_{-j})$. By construction, $p(\mathbf{x}_j \,|\, \mathbf{x}_{-j}) = p(\tilde{\mathbf{x}}_j \,|\, \mathbf{x}_j, \mathbf{y})$. Therefore, $\hat{I}(\{\boldsymbol{x}_j^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{x}_{-j}^{(i)}\}_{i=1}^N)$ must be equivalent in distribution to $\hat{I}(\{\widetilde{\boldsymbol{x}}_j^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{x}_{-j}^{(i)}\}_{i=1}^N)$. □

C.2 PROOF OF LEMMA 1

Proof. We prove that expected f-divergences are a *proper* test statistic by showing that they yield a p-value of zero under the alternate hypothesis, or a p-value which is uniformly distributed under the null. Assume we have a consistent estimator $\hat{f}$ of $\mathbb{E}_{q(\mathbf{x}_{-j})}\mathrm{D}_f\left(\cdot \,\|\, \cdot\right)$. Let $\mathcal{D}_N := \{\boldsymbol{x}_j^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{x}_{-j}^{(i)}\}_{i=1}^N$ and $\widetilde{\mathcal{D}}_{j,N} := \{\widetilde{\boldsymbol{x}}_j^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{x}_{-j}^{(i)}\}_{i=1}^N$. Recall that the p-value for our test is:

$$p_j\left(\mathcal{D}_N\right) = \mathop{\mathbb{E}}_{\widetilde{\boldsymbol{x}}_j^{(i)} \sim p(\boldsymbol{x}_j \,|\, \boldsymbol{x}_{-j}^{(i)})} \left[\mathbb{1}\left(\hat{f}\left(\mathcal{D}_N\right) \leq \hat{f}\left(\widetilde{\mathcal{D}}_{j,N}\right)\right)\right].$$

In the case where $\mathbf{x}_j \not\perp \mathbf{y} \,|\, \mathbf{x}_{-j}$, we observe that as $N \to \infty$, $\hat{f}\left(\widetilde{\mathcal{D}}_{j,N}\right) \xrightarrow{p} \mathbb{E}_{q(\mathbf{x}_{-j})}\mathrm{D}_f\left(p(\widetilde{\mathbf{x}}_j, \mathbf{y} \,|\, \mathbf{x}_{-j}) \,\|\, p(\widetilde{\mathbf{x}}_j \,|\, \mathbf{x}_{-j})p(\mathbf{y} \,|\, \mathbf{x}_{-j})\right) = 0$, where $\xrightarrow{p}$ indicates a convergence in probability. We can also state that $\hat{f}\left(\mathcal{D}_N\right) \xrightarrow{p} \mathbb{E}_{q(\mathbf{x}_{-j})}\mathrm{D}_f\left(p(\mathbf{x}_j, \mathbf{y} \,|\, \mathbf{x}_{-j}) \,\|\, p(\mathbf{x}_j \,|\, \mathbf{x}_{-j})p(\mathbf{y} \,|\, \mathbf{x}_{-j})\right) > 0$ since $\mathbf{x}_j \not\perp \mathbf{y} \,|\, \mathbf{x}_{-j}$. Therefore, the term inside the expectation above is always 0, yielding a p-value of 0 in the limit of N. Since these p-values converge in probability to a single point, this is equivalent to converging in distribution to a delta mass at 0.

Algorithm 3: k-fold AMI feature selection

Input: $\boldsymbol{x} \in \mathbb{R}^{N \times D}$, feature matrix; $\boldsymbol{y} \in \mathbb{R}^N$, labels
Output: $\mathbf{p}$, the p-values for $\boldsymbol{x}_j \perp \boldsymbol{y} \mid \boldsymbol{x}_{-j}\ \forall j$
Split $\boldsymbol{x}$ and $\boldsymbol{y}$ into M folds, $G_1, G_2, \ldots, G_M$
for $j \in [1, 2, \ldots, D]$ **do**
  Estimate $q_\theta := q_\theta(\boldsymbol{x}_j \mid \boldsymbol{x}_{-j})$
  $h \leftarrow 0$
  for $m \in [1, 2, \ldots, M]$ **do**
    Estimate $q_\beta^{(m)} := q_\beta(\boldsymbol{y}^{(G_{-m})} \mid \boldsymbol{x}^{(G_{-m})})$
    $h \leftarrow h + \frac{1}{M}\mathcal{L}_\beta(\boldsymbol{y}^{(G_m)} \mid \boldsymbol{x}^{(G_m)})$, where $\mathcal{L}_\beta$ is an log-likelihood estimate using $q_\beta^{(m)}$
  end
  Let $\tilde{\mathbf{h}}$ be a K-dimensional vector of 0s
  for $k \in [1, 2, \ldots, K]$ **do**
    Sample $\tilde{\boldsymbol{x}}_j \sim q_\theta$
    for $m \in [1, 2, \ldots, M]$ **do**
      Estimate $q_{\tilde{\beta}}^{(k,m)} := q_{\tilde{\beta}}(\boldsymbol{y}^{(G_{-m})} \mid \tilde{\boldsymbol{x}}_j^{(G_{-m})}, \boldsymbol{x}_{-j}^{(G_{-m})})$
      Let $\tilde{\mathbf{h}}^{(k)} \leftarrow \tilde{\mathbf{h}}^{(k)} + \frac{1}{M}\mathcal{L}_{\tilde{\beta}}(\boldsymbol{y}^{(G_m)} \mid \widetilde{\boldsymbol{x}}_j^{(G_m)}, \boldsymbol{x}_{-j}^{(G_m)})$, where $\mathcal{L}_{\tilde{\beta}}$ is an log-likelihood estimate using $q_{\tilde{\beta}}^{(k,m)}$
    end
  end
  Let $\mathbf{p}_j = \frac{1}{K+1}\left[1 + \sum_{k=1}^K \mathbb{1}\left(h \leq \tilde{\mathbf{h}}^{(k)}\right)\right]$
end

In the case where $\mathbf{x}_j \perp \mathbf{y} \mid \mathbf{x}_{-j}$, the distribution of $\hat{f}(\mathcal{D}_N)$, is the same as that of $\hat{f}(\widetilde{\mathcal{D}}_{j,N})$ as shown in Lemma 2.

Let $p_j^N = p_j\left(\mathcal{D}_N\right)$ and $F_N(a)$ be its CDF. Let $\rho = q(\hat{f}_N \leq a) = F_N(a)$ be the level of significance. With this, we derive the distribution of the p-value:

$$q(p_j^N \leq \rho) = q(F_N^{-1}(p_j^N) \leq F_N^{-1}(\rho)) = q(\hat{f}_N \leq a) = \rho.$$

Note that this means p_j^N is uniformly distributed under the null. This result holds regardless of N. Thus p_j^N forms a sequence of random variables, indexed by N, that are identically distributed as a uniform random variable over $[0, 1]$. This means that the sequence converges in distribution to a uniform distribution over $[0, 1]$.

This shows that expected f-divergences are proper test statistics.

□

D SIMPLIFICATION OF f-DIVERGENCE COMPARISON TO DIFFERENCE IN AMI

We show below how using the KL-divergence can reduce the number of terms to be estimated for an f-divergence. We know that using $f(z) = z \log z$ in $\mathcal{D}_f\left(\cdot||\cdot\right)$ gives us the KL-divergence which

then yields the AMI under expectation with respect to $q(x_{-j})$ (Kullback & Leibler, 1951)

$$
\begin{aligned}
&\mathbb{E}_{q(\mathbf{x}_{-j})}\mathrm{D}_f\left(q(\mathbf{x}_j, \mathbf{y} \mid \mathbf{x}_{-j}) \parallel q(\mathbf{x}_j \mid \mathbf{x}_{-j})q(\mathbf{y} \mid \mathbf{x}_{-j})\right) \\
&= \mathbb{E}_{q(\mathbf{x}_{-j})}\mathrm{KL}\left(q(\mathbf{x}_j, \mathbf{y} \mid \mathbf{x}_{-j}) \,||\, q(\mathbf{x}_j \mid \mathbf{x}_{-j})q(\mathbf{y} \mid \mathbf{x}_{-j})\right) \qquad (9) \\
&= \mathbb{E}_{q(\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j, \mathbf{y} \mid \mathbf{x}_{-j})}\left[\log \frac{q(\mathbf{x}_j, \mathbf{y} \mid \mathbf{x}_{-j})}{q(\mathbf{x}_j \mid \mathbf{x}_{-j})q(\mathbf{y} \mid \mathbf{x}_{-j})}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j, \mathbf{y} \mid \mathbf{x}_{-j})}\left[\log \frac{q(\mathbf{x}_j, \mathbf{y} \mid \mathbf{x}_{-j})}{q(\mathbf{x}_j \mid \mathbf{x}_{-j})q(\mathbf{y} \mid \mathbf{x}_{-j})}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})q(\mathbf{x}_j \mid \mathbf{x}_{-j})}\left[\log \frac{q(\mathbf{x}_j, \mathbf{y} \mid \mathbf{x}_{-j})}{q(\mathbf{x}_j \mid \mathbf{x}_{-j})q(\mathbf{y} \mid \mathbf{x}_{-j})}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})q(\mathbf{x}_j \mid \mathbf{x}_{-j})}\left[\log \frac{q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})}{q(\mathbf{y} \mid \mathbf{x}_{-j})}\right] \qquad (10) \\
&= \mathrm{AMI}(\mathbf{x}_j, \mathbf{y} \mid \mathbf{x}_{-j}) \qquad (11)
\end{aligned}
$$

We can now show the reduction of a difference in AMI terms to a difference in expected log-probabilities:

$$
\begin{aligned}
\delta_I &= \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})q(\mathbf{x}_j, \mathbf{x}_{-j})}\left[\log \frac{q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})}{q(\mathbf{y} \mid \mathbf{x}_{-j})}\right] - \mathbb{E}_{q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})q(\tilde{\mathbf{x}}_j, \mathbf{x}_{-j})}\left[\log \frac{q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})}{q(\mathbf{y} \mid \mathbf{x}_{-j})}\right] \\
&= \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})q(\mathbf{x}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})\right] - \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})q(\mathbf{x}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \mathbf{x}_{-j})\right] \\
&\quad - \mathbb{E}_{q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})q(\tilde{\mathbf{x}}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})\right] + \mathbb{E}_{q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})q(\tilde{\mathbf{x}}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \mathbf{x}_{-j})\right] \\
&= \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})q(\mathbf{x}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})\right] - \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_{-j})q(\mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \mathbf{x}_{-j})\right] \\
&\quad - \mathbb{E}_{q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})q(\tilde{\mathbf{x}}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})\right] + \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_{-j})q(\mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \mathbf{x}_{-j})\right] \\
&= \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})q(\mathbf{x}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})\right] - \mathbb{E}_{q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})q(\tilde{\mathbf{x}}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})\right] \\
&= \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})q(\mathbf{x}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})\right] - \mathbb{E}_{q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})q(\mathbf{x}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})\right] \\
&= \mathbb{E}_{q(\mathbf{x}_j, \mathbf{x}_{-j})}\left[\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})\right] - \mathbb{E}_{q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})}\left[\log q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})\right]\right] \qquad (12)
\end{aligned}
$$

Therefore, the monte-carlo estimate for Equation (12) requires estimation of only $q(\mathbf{y} \mid \mathbf{x}_j, \mathbf{x}_{-j})$, and $q(\mathbf{y} \mid \tilde{\mathbf{x}}_j, \mathbf{x}_{-j})$.

E INSTANCE-WISE FEATURE SELECTION EXAMPLES

E.1 CONSISTENT PREDICTIONS ALONE ARE INSUFFICIENT FOR INSTANCE-WISE FEATURE SELECTION

Recall our sufficiency condition for instance-wise feature selection as mentioned in Definition 3. In this example, we see what happens when this condition is not met. We notice that this definition does not suffice to reject an unimportant feature. Consider a simple data generating process:

$$
\begin{aligned}
\mathbf{y} &= \mathbf{z}\mathbf{x}_1 + (1 - \mathbf{z})\mathbf{x}_2 + \epsilon \\
\mathbf{z} &\sim \text{Bernoulli}(0.5) \\
\mathbf{x}_1, \mathbf{x}_2 &\sim \mathcal{N}(0, \sigma_{\mathbf{x}}^2) \\
\epsilon &\sim \mathcal{N}(0, \sigma_{\epsilon}^2)
\end{aligned}
$$

where $\boldsymbol{z}$ is not observed. We can now write out the probability distributions we care about. Note that taking an expectation like $\mathbb{E}_{\widetilde{\mathbf{x}}_1 \sim q(\mathbf{x}_1 \mid \mathbf{x}_2)}q(\mathbf{y} \mid \widetilde{\mathbf{x}}_1\mathbf{x}_2)$ yields $q(\mathbf{y} \mid \mathbf{x}_2)$. For simplicity, we leave out

the use of complete conditions and work directly with the latter probability distributions:

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) &= \mathcal{N}\left(\mathbf{z}\mathbf{x}_1 + (1-\mathbf{z})\mathbf{x}_2, \sigma_\epsilon^2\right) \\
p(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2) &= \int p(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) p(\mathbf{z}|\mathbf{x}_1, \mathbf{x}_2) dz = \int p(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) p(\mathbf{z}) dz \\
&= \frac{1}{2}\mathcal{N}\left(\mathbf{x}_1, \sigma_\epsilon^2\right) + \frac{1}{2}\mathcal{N}\left(\mathbf{x}_2, \sigma_\epsilon^2\right) \\
p(\mathbf{y}|\mathbf{x}_2) &= \frac{1}{2}\mathcal{N}\left(0, \sigma_{\mathbf{x}}^2 + \sigma_\epsilon^2\right) + \frac{1}{2}\mathcal{N}\left(\mathbf{x}_2, \sigma_\epsilon^2\right) \\
p(\mathbf{y}|\mathbf{x}_1) &= \frac{1}{2}\mathcal{N}\left(\mathbf{x}_1, \sigma_\epsilon^2\right) + \frac{1}{2}\mathcal{N}\left(0, \sigma_{\mathbf{x}}^2 + \sigma_\epsilon^2\right)
\end{aligned}
$$

Now consider an instance $(\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{z}^{(i)})$ where $\boldsymbol{z}^{(i)} = 1$ which means that $\boldsymbol{y}^{(i)}$ depends only on feature $\boldsymbol{x}_1^{(i)}$. Now we check if $p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}) > p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)})$. Using our definitions from earlier, we can expand this inequality:

$$
p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}) - p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}) = \frac{1}{2}\mathcal{N}\left(\boldsymbol{y}^{(i)}; \boldsymbol{x}_2^{(i)}, \sigma_\epsilon^2\right) - \frac{1}{2}\mathcal{N}\left(\boldsymbol{y}^{(i)}; 0, \sigma_{\boldsymbol{x}}^2 + \sigma_\epsilon^2\right)
$$

For all i such that $\boldsymbol{y}^{(i)}$ lies in a non-0 interval around $\boldsymbol{x}_2^{(i)}$, we have that $p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}) - p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}) > 0$. For example let $\sigma_\epsilon = \sigma_x = 1$, then $\boldsymbol{x}_2^{(i)} = 5$, we have that $\boldsymbol{y} \in [3, 17]$ satisfies this. This means that x_2 will be deemed important as per the candidate Definition 3.

E.2 Instance-wise score example

In this example, we see how scores computed using Equation (8) can help identify important features for a given instance, under the assumptions stated in Definition 3. Consider a simple data generating process:

$$
\begin{aligned}
\mathbf{y} &= \mathbf{z}\mathbf{x}_1 + (1-\mathbf{z})\mathbf{x}_2 \\
\mathbf{z}, \mathbf{x}_1, \mathbf{x}_2 &\sim \text{Bernoulli}(0.5)
\end{aligned}
$$

where all random variables are observed. Let us now consider the following observed instance: $(\boldsymbol{y}^{(i)}, \boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \boldsymbol{z}^{(i)}) = (1, 0, 1, 0)$. We can now devise a test for each of $\boldsymbol{x}_1^{(i)}$, $\boldsymbol{x}_2^{(i)}$, and $\boldsymbol{z}^{(i)}$. For $\boldsymbol{x}_1^{(i)}$, we want to check:

$$
p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \boldsymbol{z}^{(i)}) > p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_2^{(i)}, \boldsymbol{z}^{(i)}) \tag{13}
$$

We can create similar tests for the other two variables as well:

$$
p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \boldsymbol{z}^{(i)}) > p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}, \boldsymbol{z}^{(i)}) \tag{14}
$$

$$
p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \boldsymbol{z}^{(i)}) > p(\boldsymbol{y}^{(i)}|\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}) \tag{15}
$$

We can use Table 3 to help evaluate Equations (13) to (15):

$$
\begin{aligned}
p(\boldsymbol{y}^{(i)} = 1|\boldsymbol{x}_1^{(i)} = 0, \boldsymbol{x}_2^{(i)} = 1, \boldsymbol{z}^{(i)} = 0) &= 1 \\
p(\boldsymbol{y}^{(i)} = 1|\boldsymbol{x}_2^{(i)} = 1, \boldsymbol{z}^{(i)} = 0) &= 1 \\
p(\boldsymbol{y}^{(i)} = 1|\boldsymbol{x}_1^{(i)} = 0, \boldsymbol{z}^{(i)} = 0) &= 0.5 \\
p(\boldsymbol{y}^{(i)} = 1|\boldsymbol{x}_1^{(i)} = 0, \boldsymbol{x}_2^{(i)} = 1) &= 0.5
\end{aligned}
$$

meaning $\boldsymbol{x}_1^{(i)}$ is not important to $\boldsymbol{y}^{(i)}$, but $\boldsymbol{x}_2^{(i)}$ and $\boldsymbol{z}^{(i)}$ are important.

F Simulated data feature selection - additional results

In this section, we present additional results that use AMI as a test statistic in a HRT framework. This offers a significant speedup as the HRT framework avoids having to refit estimators using CCKs. Figure 4 shows a quantile-quantile plot of the null p-values for each FDR-controlling feature selection method. We notice that both HRT-based methods tend to deflate p-values. This often results in features being mistakenly selected as important. Using the same test statistic, AMI, in a CRT helps mitigate this issue.

x_1	x_2	z	y
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

Table 3: Full distribution for example in Appendix E.2

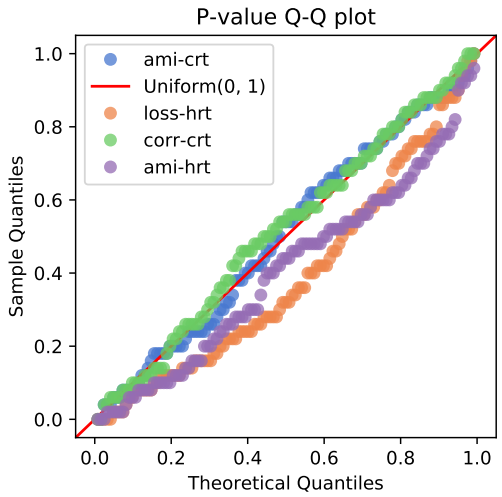

Figure 4: Quantile-Quantile plot showing uniformity of p-values across various FDR-controlling methods.

G SELECTOR INSTANCE-WISE FEATURE SELECTION

Table 4 shows the percentage of important features identified by each baseline feature selection algorithm. Table 5 shows the percentage of samples where the selector variable $x_1^{(i)}$ was identified as important. We notice that AMI-CRT outperforms baselines quite significantly in detecting the selector feature.

	ami-crt	loss-hrt	corr-crt	lime	shap	rf
selector	**0.67**	0.45	0.33	0.57	0.61	0.33

Table 4: Percentage of important features identified by various feature selection algorithms. In each instance, only 6 features are important. The score for each method is an average of the percentages of important features identified across each instance. Each method is allowed to pick only 6 features. We see that `ami-crt` selects more important features than state-of-the-art baselines.

H CELIAC DISEASE GENOMIC FEATURE SELECTION

Table 7 shows the set of SNPs deemed significant by AMI-CRT. We annotate each SNP with its position in the human genome, and whether it was previously reported as significant in a biological study.

	ami-crt	loss-hrt	corr-crt	lime	shap	rf
selector	**0.97**	0.85	0.33	0.25	0.37	0.33

Table 5: Percentage of instances where the selector variable $\widetilde{\boldsymbol{x}}_1^{(i)}$ was identified by various feature selection algorithms. This is the only feature that is relevant to all samples. We observe that `ami-crt` almost always identifies the selector feature ($\mathbf{x}_1$) as important.

Method	Significant features returned	Percentage biologically significant
`ami-crt`	17	**76.47%**
`corr-crt`	*185*	6.40%
`loss-hrt`	14	57.14%

Table 6: The number of significant features reported at a 20% FDR level for each test, and the percentage of features previously identified in a biological study.

I HOSPITAL READMISSION FEATURE SELECTION

For the hospital readmission dataset, we applied several standard pre-processing techniques. First, we binarized the labels so that a label of 1 indicates a readmission event, and a label of 0 indicates no readmission event. We then encoded each categorical feature as a one-hot encoding. We then imputed missing values using the median across the dataset, and dropped the "weight" feature as it was found to be 97% missing.

To sample from the complete conditional distributions $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$, we used a neural network to fit the complete conditional regression detailed in Equation (6). For continuous values of $\mathbf{x}_j$, we first discretized the data into bins, then used our neural network to predict the bins. To map the bins back to values in the domain of $\mathbf{x}_j$, we used the mean of the range of values in each bin.

J IMAGENET INSTANCE-WISE FEATURE SELECTION

Figure 5 shows some of the results of instance-wise feature selection on ImageNet data using AMI-CRT. Figures 6 and 7 show results on the same task, using LIME and SHAP respectively. We notice that AMI-CRT identifies patches that seem more likely to help differentiate between ambulances and policevans. AMI-CRT identifies relevant text like the words "ambulance" or "police" that are very likely to help distinguish between the two classes. LIME identifies some relevant features of the image like wheels and lights, but fails to identify relevant words. SHAP does a good job at identifying distinguishing symbols like the caduceus and the FBI logo, but occasionally misses out on relevant text.

Position	SNP	Featured in previous study
chr2:102454108	rs917997	yes
chr2:68371823	rs17035378	yes
chr3:159947262	rs17810546	yes
chr3:188394766	rs1464510	yes
chr3:46193709	rs13098911	yes
chr4:122194347	rs13151961	yes
chr6:137651931	rs2327832	yes
chr6:26451325	rs2237236	yes
chr6:28423688	rs2859365	no
chr6:29505139	rs757256	no
chr6:29844253	rs2734994	no
chr6:31642909	rs1052486	no
chr6:32638107	rs2187668	yes
chr6:90216893	rs10806425	yes
chr11:128511079	rs11221332	yes
chr12:111569952	rs653178	yes
chr21:44227538	rs4819388	yes

Table 7: This table details each SNP returned by AMI-CRT, whether it was featured in a previous biological study relating to Celiac disease, and its position on the human genome.

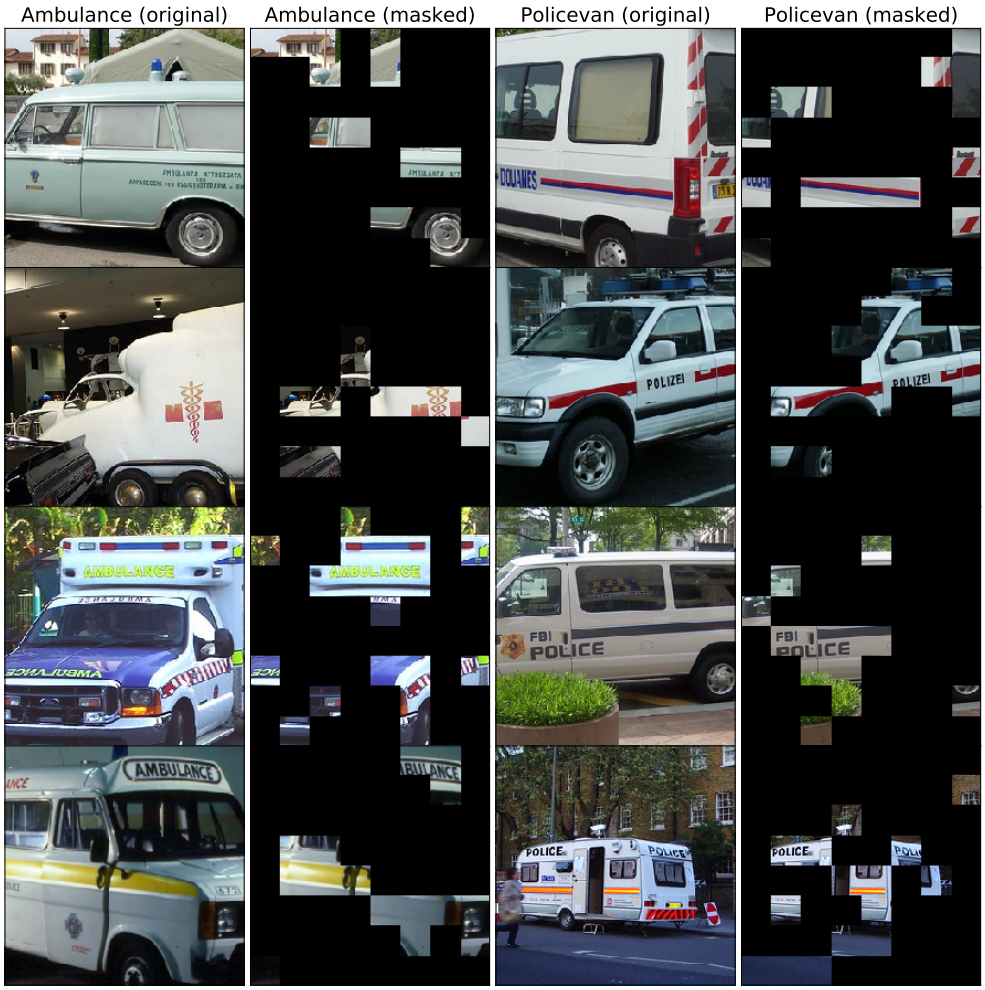

Figure 5: Instance-wise feature selection using AMI-CRT. The first and third columns show the original image of ambulances or policevans respectively. The second and fourth columns show only the patches which were found to have non-zero AMI with the label, given the rest of the patches.

Figure 6: Instance-wise feature selection using LIME. The first and third columns show the original image of ambulances or policevans respectively. The second and fourth columns show only the patches which were found to be important.

Figure 7: Instance-wise feature selection using SHAP. The first and third columns show the original image of ambulances or policevans respectively. The second and fourth columns show patches that are found to contribute to the label. Green indicates a patch found relevant for the ambulance class, and red indicates a patch found relevant for the policevan class.