Covered Information Disentanglement: Model Transparency via Unbiased Permutation Importance

João P. B. Pereira[1,2], **Erik S.G. Stroes**[1], **Aeilko H. Zwinderman**[1], **Evgeni Levin**[1,2]

[1]Amsterdam University Medical Center, Meibergdreef 9 1105 AZ, Amsterdam, The Netherlands
[2] Horaizon, Marshallaan 2 2625 GZ, Delft, The Netherlands
{j.p.belopereira, e.levin}@amsterdamumc.nl

Abstract

Model transparency is a prerequisite in many domains and an increasingly popular area in machine learning research. In the medical domain, for instance, unveiling the mechanisms behind a disease often has higher priority than the diagnostic itself since it might dictate or guide potential treatments and research directions. One of the most popular approaches to explain model global predictions is the *permutation importance* where the performance on permuted data is benchmarked against the baseline. However, this method and other related approaches will undervalue the importance of a feature in the presence of covariates since these cover part of its provided information. To address this issue, we propose Covered Information Disentanglement (*CID*), a framework that considers all feature information overlap to correct the values provided by *permutation importance*. We further show how to compute *CID* efficiently when coupled with *Markov random fields*. We demonstrate its efficacy in adjusting *permutation importance* first on a controlled toy dataset and discuss its effect on real-world medical data.

Introduction

Understanding the biological underpinnings of disease is at the core of medical research. Model transparency and feature relevance are thus a top priority to discover new potential treatments or research directions. One of the current most popular methods to explain local model predictions is *SHAP* (Lipovetsky and Conklin 2001; Štrumbelj and Kononenko 2014; Lundberg and Lee 2017), a game-theoretic approach that considers the features as "players" and measures their marginal contributions to all possible feature subset combinations. *SHAP* has also been generalized in *SAGE* (Covert, Lundberg, and Lee 2020) to compute global feature importance. However, recent work by Kumar et al. (Kumar et al. 2020) exposes some mathematical issues with *SHAP* and concludes that this framework is ill-suited as a general solution to quantifying feature importance. Other local-based methods such as *LIME* (Ribeiro, Singh, and Guestrin 2016) and its variants (see e.g. (Singh, Ribeiro, and Guestrin 2016; Ribeiro, Singh, and Guestrin 2018.; Guidotti et al. 2018; Pereira et al. 2019)) build weak yet explainable models on the neighborhood of each instance. While this achieves higher prediction transparency for each data point, in this work, we are mainly concerned with a more holistic view of importance, which may be more appropriate to guide new research directions and unravel disease mechanisms. Tree-based methods are very commonly selected for this purpose because they compute the impurity or *Gini importance* (Breiman 2001). The impurity importance is biased in favor of variables with many possible split points; i.e. categorical variables with many categories or continuous variables (Strobl et al. 2007). A generally accepted alternative to computing the *Gini importance* is the *permutation importance* (Breiman 2001), which benchmarks the baseline performance against permuted data. There is, however, the issue of multicollinearity. When features are highly correlated, feature permutation will underestimate the individual importance of at least one of the features, since a great deal of the information provided by this feature is "covered" by its covariates. One option is to permute correlated features together (Toloşi and Lengauer 2011). However, this implies choosing an arbitrary correlation grouping threshold. Most importantly, it misses the differentiation between each feature's contribution to the final prediction. Motivated by the idea that there is an information overlap between different features, we develop Covered Information Disentanglement (*CID*),[1] an information-theoretic approach to disentangle the shared information and scale the *permutation importance* values accordingly. We demonstrate how *CID* can recover the right importance ranking on artificial data and discuss its efficacy on the Cardiovascular Risk Prediction dataset (Hoogeveen et al. 2020).

Methodology

Notation We denote matrices, 1-dimensional arrays, and scalars/functions with capital bold, bold, and regular text, respectively (e.g. $\mathbf{X}$, $\mathbf{x}$, α/f). Given a dataset $\mathbf{X}_{M\times N}$, we will denote its random variables by capital regular text with a subscript and the values using lowercase (e.g. X_i and x_i), while the joint density/mass will be represented as $p(x)$. The expected loss of a function given by: $\frac{1}{M}\sum_{i=1}^{M} l\left[y,\, f(\mathbf{x}_i)\right]$

[1]We make an implementation of *CID* publicly available at: https://github.com/JBPereira/CID.

will be denoted by $\mathcal{L}\left[f\left(\mathbf{X}\right)\right]$.

Information Theory Background

Information theory (IT) is a useful tool used in quantifying relations between random variables. The basic building block in IT is the *entropy* of an r.v. X_i, which is defined as: $H(X_i) \equiv -\sum_{x_i} p(x_i) \log p(x_i)$. The *joint entropy* between r.v.s X_i and X_j is defined as: $H(X_i, X_j) \equiv -\sum_{x_i} \sum_{x_j} p(x_i, x_j) \log p(x_i, x_j)$. The *mutual information* between r.v.s X_i and X_j is the relative entropy between the joint entropy and the product distribution $p(x_i)p(x_j)$: $I(X_i, X_j) \equiv \sum_{x_i} \sum_{x_j} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)}$. For a more thorough exposition to IT, the reader can refer to (Cover and Thomas 2012).

Using the definitions above, one can derive properties that resemble those of set theory, where joint entropy and mutual information are the information-theoretic counterparts to union and intersection, respectively (Ting 2008). In order to keep this intuition when generalizing to higher dimensions, one can define the entropy of the union of N features as:

Definition 1. ***Multivariate Union Entropy***

$$H\left(\cup_{i=1}^{N} X_i\right) \equiv -\sum_{x_i} p(x_1, ..., x_N) log\, p(x_1, ..., x_N)$$

and using the Inclusion-Exclusion principle, we can define the intersection as:

Definition 2. ***Multivariate Intersection Entropy***

$$H\left(\cap_{i=1}^{N} X_i\right) \equiv \sum_{x_1,...,x_N} p(x_1, ..., x_N) h_{ci}(x_1, ..., x_N),$$

$$h_{ci}(x_1, ..., x_N) = \sum_{k=1}^{N} (-1)^{k-1} \sum_{\substack{I \subseteq \{1, ..., N\}; \\ |I|=k}} h(x_{I_1}, ..., x_{I_k}),$$

$h(\mathbf{x}) = -log\, p(\mathbf{x})$ *and* h_{ci} *is the local co-information.*

This definition of multivariate intersection is also called co-information and it may yield negative values. This can happen for instance if X_i has no correlation with X_I but knowing X_I introduces a correlation between the two (what is commonly known as 'explaining away'). This motivated Williams and Beer to draw the distinction between redundant and synergistic information and propose *partial information decomposition* (PID) (Williams and Beer 2010). Ince (Ince 2017) thoroughly analyzed the multivariate properties of PID directly applied to multivariate entropy and suggested to divide the individual terms in definition 2, so that positive local entropy terms correspond to redundant entropy, while the negative ones correspond to synergistic entropy.

Permutation Feature Importance

Feature importance is a subjective notion that may vary with application. Consider a supervised learning task where a model f is trained/tested on dataset $\mathbf{X}, \mathbf{y}$ and its performance is measured by a function $\mathcal{L}$. In this work, we will refer to feature importance as the extent to which a feature X_i affects $\mathcal{L}[f(\mathbf{X})]$, on its own and through its interactions with $X_{\setminus\{i\}}$. Permutation importance was first introduced by Breiman (Breiman 2001) in random forests as a way to understand the interaction of variables that is providing the predictive accuracy.

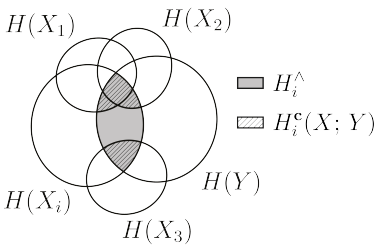

Figure 1: An illustration of the *permutation importance* bias in the presence of covariates and the measures needed to correct it. The mutual information between random variable X_i and Y (represented in gray) is covered by the information provided by r.v.s X_1, X_2 and X_3. Permutation importance only measures the non-covered part (non-shaded gray), and to correct its value, we suggest computing $H_i^{\mathbf{c}}(X;\, Y)$.

Consider a dataset $\mathbf{X}_{M \times N}$ and denote the jth instance of the ith feature by $\mathbf{X}_i^j$. Suppose the set $\{1, ..., M\}$ is sampled and denote the subsample by $\mathbf{s}$, $\mathbf{s} \subseteq \{1, ..., M\}$. Consider further a random permutation of this subset which we denote by $\boldsymbol{\pi}(\mathbf{s})$ and its jth element by $\boldsymbol{\pi}_j(\mathbf{s})$. The *permutation importance*, is given by:

$$e_i(f, \mathbf{s}) = \sum_{j \in \mathbf{s}}^{|\mathbf{s}|} \Bigg(\mathbf{E}_{\sim p(\boldsymbol{\pi})} \left[\mathcal{L}\left(f\left(\mathbf{X}_1^j, ..., \mathbf{X}_i^{\boldsymbol{\pi}_j(\mathbf{s})}, ..., \mathbf{X}_N^j \right) \right) \right] - \mathcal{L}\left(f\left(\mathbf{X}_1^j, ..., \mathbf{X}_N^j \right) \right) \Bigg) \quad (1)$$

$$e_i(f) = \mathbf{E}_{\sim p(\mathbf{s})} \left[e_i(f, \mathbf{s}) \right] \quad (2)$$

Covered Information Disentanglement

In the presence of covariates, the *permutation importance* measures the performance dip caused by removing the non-mutual information between the feature and the remaining data. That is:

$$e_i(f) = \mathcal{I}_i(f) - e_i^{\cup}(f), \quad (3)$$

where $\mathcal{I}_i(f) = \mathbf{E}_{\sim p(\mathbf{s})} \left[\mathcal{I}_i(f, \mathbf{s}) \right]$ is the expected total importance of feature i under model f (the quantity we are interested in) and $e_i^{\cup}(f) = \mathbf{E}_{\sim p(\mathbf{s})} \left[e_i^{\cup}(f, \mathbf{s}) \right]$ is the expected performance dip covered by all other variables. To compute $e_i^{\cup}(f)$ would require applying the Inclusion-Exclusion principle and measuring the performance dip for all possible feature combinations of size 1 to the number of features. Instead, we note that $e_i^{\cup}(f)$ intuitively measures the model performance dip when the model is deprived of the information covered by the r.v.s that are correlated with X_i. For an intuitive depiction of the problem, see figure 1.

Motivated by the analogy between set-theory and information measures, we define the joint information

between an r.v. and the target variable that is "covered" by the other r.v.s as:

Definition 3. ***Covered information (CI)*** *Given an r.v. X_i and a set of distinct r.v.s $X_{\mathbf{i}^-}$, $\mathbf{i}^- = \{1, ...N\}\backslash\{i\}$, the information of X_i w.r.t. Y covered by $X_{\mathbf{i}^-}$ is defined as:*

$$H_i^{\mathbf{c}}(X;\, Y) = H\left(X_i \cap Y \cap \left\{\cup_{j\in\mathbf{i}^-} X_j\right\}\right).$$

When it is clear from the context what Y and $X_{\mathbf{i}^-}$ are, we will abbreviate $H_i^{\mathbf{c}}(X;\, Y)$ into $H_i^{\mathbf{c}}$, denote the mutual information with Y by $H_i^{\wedge}$, and the respective local co-information terms for the kth row in the dataset with $h_{ik}^{\mathbf{c}} \equiv h_i^{\mathbf{c}}(\mathbf{X}_i^k, Y^k)$ and $h_{ik}^{\wedge} \equiv h_i^{\wedge}(\mathbf{X}_i^k, \mathbf{y}^k)$. We further divide $H_i^{\mathbf{c}}$ and $H_i^{\wedge}$ into its redundant and synergistic counterparts, which for a specific sample $\mathbf{s}$ are given by:

$$\textbf{Redundant MI}: H_i^{\wedge^+}(\mathbf{s}) = \frac{1}{|\mathbf{s}|}\sum_{k\in\mathbf{s}} \max\left(0, h_{ik}^{\wedge}\right)$$

$$\textbf{Synergistic MI}: H_i^{\wedge^-}(\mathbf{s}) = \frac{1}{|\mathbf{s}|}\sum_{k\in\mathbf{s}} \left|\min\left(0, h_{ik}^{\wedge}\right)\right|$$

$$\textbf{Redundant CI}: H_i^{\mathbf{c}^+}(\mathbf{s}) = \frac{1}{|\mathbf{s}|}\sum_{k\in\mathbf{s}} \max\left(0, h_{ik}^{\mathbf{c}}\right)$$

$$\textbf{Synergistic CI}: H_i^{\mathbf{c}^-}(\mathbf{s}) = \frac{1}{|\mathbf{s}|}\sum_{k\in\mathbf{s}} \left|\min\left(0, h_{ik}^{\mathbf{c}}\right)\right|$$

Assumption 1. *Permutation importance and entropy terms are related through a map $\phi_f : \mathbb{R}^4 \to \mathbb{R}$, such that $e_i(f, \mathbf{s}) = \phi_f\left(H_i^{\mathbf{c}^+}(\mathbf{s}), H_i^{\mathbf{c}^-}(\mathbf{s}), H_i^{\wedge^+}(\mathbf{s}), H_i^{\wedge^-}(\mathbf{s})\right) + \epsilon$, where ϵ is an error term.*

Thus, if assumption 1 holds, we can use the information of X_i w.r.t. Y by $X_{\mathbf{i}^-}$ and approximate equation 3 with:

$$e_i^{\cup}(f, \mathbf{s}) \approx \phi_f\left(0, H_i^{\mathbf{c}^-}(\mathbf{s}), H_i^{\wedge^+}(\mathbf{s}), H_i^{\wedge^-}(\mathbf{s})\right) - \phi_f\left(H_i^{\mathbf{c}^+}(\mathbf{s}), H_i^{\mathbf{c}^-}(\mathbf{s}), H_i^{\wedge^+}(\mathbf{s}), H_i^{\wedge^-}(\mathbf{s})\right). \quad (4)$$

This means we can approximate the result of permuting all possible combinations of features by computing only the single-feature permutation loss and the covered information of r.v. X_i by all the others. Here, we are implicitly defining: $\mathcal{I}_i(f, \mathbf{s}) \equiv \phi_f\left(0, H_i^{\mathbf{c}^-}(\mathbf{s}), H_i^{\wedge^+}(\mathbf{s}), H_i^{\wedge^-}(\mathbf{s})\right)$, and thus the true importance in the performance difference scale is given by mapping the entropy values when there is no redundant entropy to the space of performance differences.

Since we are predicting the feature importance using a map between entropy terms (which measure model-agnostic importance) and *permutation importance* values, the end result depends only on how learnable is the model behavior w.r.t to entropy. Moreover, since the entropy values are computed for the different subsample sets $\mathbf{s}$, the overall importance variability is also estimated.

For two datasets where $I(X_i, Y) > I(X_i, Y')$ but the covered info of $(X_i, Y) > (X_i, Y')$, ***CID*** would correctly value $\mathcal{I}_i(f) > \mathcal{I}_i(f')$ which is not guaranteed using Shapley based methods since the contributions to subsets of features correlated with X_i are biased. The Shapley efficiency+symmetry properties also imply that correlated features' scores are scaled down. To see this, consider $X_i = X_j$, then symmetry$\to$ $\phi_i(v_f) = \phi_j(v_f)$ and efficiency$\to$ $\phi_i(v_f) = \phi_j(v_f) = (v_f(D) - \sum_{k\neq i,j}\phi_k(v_f))/2$. In contrast, ***CID*** values do not sum to the complete data performance, but rather are meaningful individually.

There is still the issue of computing $H_i^{\mathbf{c}}$, since it involves computing $p(X)$. Since directionality is irrelevant for the purpose of computing overlapping information, we suggest to model $p(X)$ using an undirected graphical model (UGM). Let $G = (V, E)$ denote a graph with N nodes, corresponding to the $\{X_1, ..., X_N\}$ features, and let $\mathcal{C}$ be a set of cliques (fully-connected subgraphs) of the graph G. Denoting a set of clique-potential functions by $\{\psi_{\mathcal{C}} : \mathcal{X}^{|\mathcal{C}|} \to \mathbb{R}\}$, the distribution of a *Markov random field* (***MRF***) (Koller and Friedman 2009) is given by: $p(x) = \prod_{c\in\mathcal{C}}\psi_c(x_c)/\mathbf{Z}$, where $\mathbf{Z} = \int\prod_{c\in\mathcal{C}}\psi_c(x_c)dx$ is the partition function. By the Hammersley-Clifford theorem, any distribution that can be represented in this way satisfies: $X_i \perp X_j | X_{\mathcal{N}(X_i)}$ for any $X_j \notin \mathcal{N}(X_i)$, where $\mathcal{N}(X_i)$ is the set $\{X_k : (i,k) \in E\}$. This allows to significantly simplify the expression of covered information yielding the main result of this paper:

Theorem 1. *Consider an r.v. X_i and set of r.v.s $X_{\mathbf{i}^-}$, $\mathbf{i}^- = \{1, ..., N\}\backslash\{i\}$, a response r.v. Y, as well as the set of r.v.s that are neighbors to both X_i and Y: $X_{\mathcal{N}(i,y)}$, $\mathcal{N}(i, y) \in \cup\{\mathcal{N}(X_i), \mathcal{N}(Y)\}$. For a Markov random field, the covered information of X_i by $X_{\mathbf{i}^-}$ w.r.t. Y is given by:*

$$H_i^{\mathbf{c}} = H_i^{\wedge} - \mathbf{E}_{\sim p(x_{\mathcal{N}(i,y)})}\left[log\left(f\frac{\mathbf{d}^T\mathbf{F}\,\mathbf{e}}{\mathbf{d}^T\mathbf{F}_y\mathbf{F}_{x_i}^T\mathbf{e}}\right)\right],$$

where $\mathbf{F}$ is a matrix with the product of joint potential values $\psi_{\mathcal{C}_F}$ for set of cliques $F : \{c \,|\, X_i, Y \in c\}$; f, $\mathbf{F}_y$ and $\mathbf{F}_{x_i}$ are an entry, column, and row of $\mathbf{F}$, respectively, while $\mathbf{d}$ and $\mathbf{e}$ are arrays with the product of potential values $\psi_{\mathcal{C}_D}$, $\psi_{\mathcal{C}_E}$ for set of cliques $D : \{c \,|\, X_i \in c, Y \notin c\}$ and $E : \{c \,|\, X_i \notin c, Y \in c\}$ with fixed $X_{\mathbf{i}^-}$.

Proof. Using definition 1, 2 and 3:

$$H_i^{\mathbf{c}} = H_i^{\wedge} + \overbrace{H(X_i \cup Y \cup X_{\mathbf{i}^-})}^{①} - \overbrace{H(X_{\mathbf{i}^-} \cup Y)}^{②} + \overbrace{H(X_{\mathbf{i}^-})}^{③} - \overbrace{H(X_i \cup X_{\mathbf{i}^-})}^{④}.$$

The probability density for Markov Random fields is equal to $p(x) = \prod_{c\in\mathcal{C}}\psi_c(x_c)/\mathbf{Z}$, where $\mathbf{Z}$ is the partition function and $\mathcal{C}$ is the set of cliques in the Markov network. Define two sets of cliques: $A : \{c \,|\, X_i \in c\}$ and $B : \{c \,|\, X_i \notin c\}$. In that case (ignoring the partition function term because it cancels out):

$$① = -\sum_x p(x)\left[log\prod_{b\in B}\psi_b(x_b) + log\prod_{a\in A}\psi_a(x_a)\right],$$

$$② = -\sum_x p(x)\left[log\prod_{b\in B}\psi_b(x_b) + log\sum_{x_i}\prod_{a\in A}\psi_a(x_a)\right],$$

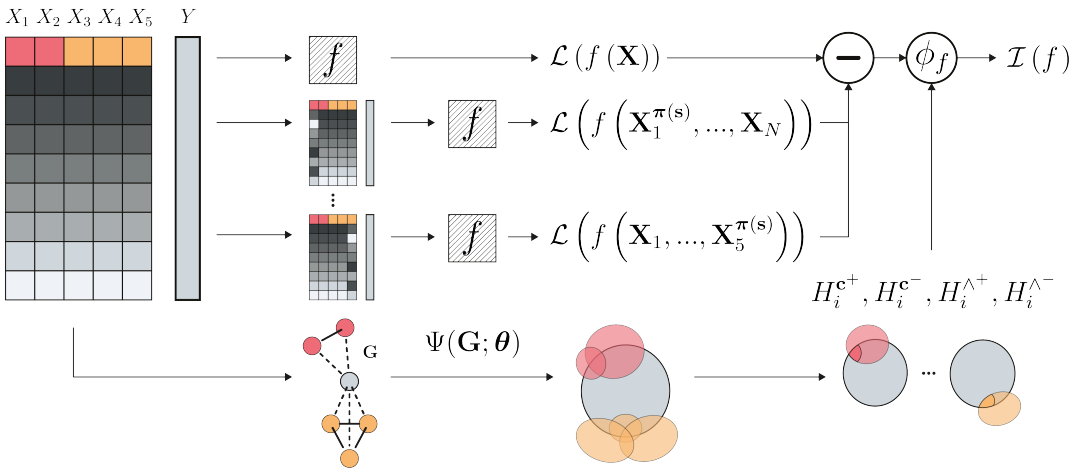

Figure 2: *CID* importance diagram. The permutation feature importance is computed by first calculating the expected loss of the model f ($\mathcal{L}(f(\mathbf{X}))$). Then, each feature's values are permuted and the expected loss of f computed. Subtracting each permuted dataset loss to the original one yields the *permutation importance*. *CID* starts by inferring the network $\mathbf{G}$ for the *Markov random field* Ψ (alternatively, a prior network is given), then the *MRF* parameters $\boldsymbol{\theta}$ are inferred, and finally, $H_i^c/H_i^\wedge$ are computed for each feature, which are then used to train the entropy/PI model ϕ_f and predict the true importance $\mathcal{I}(f)$.

$$① - ② = -\sum_x p(x) log\left(\frac{\prod_{a\in A}\psi_a(x_a)}{\sum_{x_i}\prod_{a\in A}\psi_a(x_a)}\right).$$

To compute ③ − ④, define four sets of cliques: $C : \{c \,|\, X_i \notin c, Y \notin c\}$, $D : \{c \,|\, X_i \in c, Y \notin c\}$, $E : \{c \,|\, X_i \notin c, Y \in c\}$, and $F : \{c \,|\, X_i \in c, Y \in c\}$. In order to reduce the clutter, we will introduce the following functions: $d(x_i, x_{\mathbf{i}^-}) = \prod_{j\in\mathbf{i}^-, j\sim i}\psi(x_i, x_j)$, $e(y, x_{\mathbf{i}^-}) = \prod_{j\in\mathbf{i}^-, j\sim y}\psi(y, x_j)$, $f(x_i, y) = \psi(x_i, y)$, where we will abbreviate $d(x_i, x_{\mathbf{i}^-})$ into $d(x_i)$ and $e(y, x_{\mathbf{i}^-})$ into $e(y)$ when the value for random variable $X_{\mathbf{i}^-}$ is fixed. Then (again, ignoring the partition function):

$$③ = -\sum_x p(x)\left[log\prod_{c\in C}\psi_c(x_c) + log\sum_{x_i}\sum_y d(x_i)e(y)f(x_i, y)\right],$$
$$④ = -\sum_x p(x)\left[log\prod_{c\in C}\psi_c(x_c) + log\sum_y d(x_i)e(y)f(x_i, y)\right],$$
$$③ - ④ = -\sum_x p(x) log\left(\frac{\sum_{x_i}\sum_y d(x_i)e(y)f(x_i, y)}{\sum_y d(x_i)e(y)f(x_i = X_i, y)}\right),$$

where $f(x_i = X_i, y)$ is the function f for a fixed value of the r.v. X_i. Since the set of cliques $A = \{D \cup F\}$, and denoting by $d(X_i)$, $f(X_i, Y)$ the functions d and f for fixed values of X_i and Y, then:

$$(① - ②) + (③ - ④) =$$
$$-\sum_x p(x) log\left(\frac{\sum_{x_i}\sum_y d(X_i)d(x_i)f(X_i,Y)e(y)f(x_i,y)}{\sum_{x_i}\sum_y d(X_i)d(x_i)f(x_i,Y)e(y)f(X_i,y)}\right)$$
$$= -\mathbf{E}_{\sim p(x_{\mathcal{N}(i,y)})}\left[log\, f(X_i, Y) + log\left(\frac{\mathbf{d}^T\mathbf{F}\,\mathbf{e}}{\mathbf{d}^T\mathbf{F}_y\mathbf{F}_{x_i}\mathbf{e}}\right)\right],$$

where $x_{\mathcal{N}(i,y)}$ is an instance of the set of r.v.s that are neighbors to either X_i or Y, $\mathbf{d}$ and $\mathbf{e}$ are column arrays with the different values of $d(x_i)$ and $e(y)$ for fixed $X_{\mathbf{i}^-}$, $\mathbf{F}$ is a matrix with all the values $f(x_i, y)$ with varying values of X_i in the rows and Y in the columns, and $\mathbf{F}_y$ and $\mathbf{F}_{x_i}$ are row and column vectors of $\mathbf{F}$ corresponding to fixed Y and fixed X_i, respectively. This yields the result of the theorem. □

Considerations and Simplifications If a 2-clique *MRF* is chosen, then $\mathbf{F}$ depends only on X_i and Y, and can be computed before the expectation.

***Gaussian MRF*:** Learning an *MRF*'s network structure is expensive. One popular approach is to use *graphical lasso* (Friedman, Hastie, and Tibshirani 2008) which learns the entries of a Gaussian precision matrix by finding: $\min_{\boldsymbol{\Lambda}\in\mathbb{S}_+^n} -log\ det(\boldsymbol{\Lambda}) + tr(\mathbf{S}\boldsymbol{\Lambda}) + \rho||\boldsymbol{\Lambda}||_1$, where $\boldsymbol{\Lambda}$ is the precision matrix (constrained to belong to $\mathbb{S}_+^n$, the set of positive semi-definite $n\times n$ matrices), $\mathbf{S}$ is the empirical covariance matrix and ρ acts in analogy to Lasso regularization by penalizing a large number of non-zero precision entries. We can model the potentials using *Gaussian Markov random field*s whose potentials are $\psi_{s,t}(x_s, x_t) = \exp\left[-\frac{1}{2}x_s\Lambda_{st}x_t\right]$, $\psi_s(x_s) = \exp\left[-\frac{1}{2}\left(x_s^2\Lambda_{ss} - 2\eta_s x_s\right)\right]$, where $\boldsymbol{\eta} = \boldsymbol{\Lambda}\boldsymbol{\mu}$ ($\boldsymbol{\mu}$ is the mean vector).

Discrete Approximation: Continuous *MRF* such as Gaussian Markov Random fields depend on a continuous multivariate distribution and thus the entropy must be replaced by differential entropy, which violates many of the desired properties of discrete entropy. Therefore, we will approximate a continuous distribution with a discrete one $p(x_i) \approx \delta_i p(\overline{x_i})$, where δ_i is the ith feature bin size and $\overline{x_i}$ is the mean value of the bin, and then carry on with our computations as specified in theorem 1. For the case where all bins have the same size per feature, all the δs cancel out.

Complexity: If we approximate the expectation in theorem 1 with the empirical expectation, then the asymptotic complexity becomes $\mathcal{O}(SB^2)$, where S is the number of

samples and B is the maximum between the number of bins used to discretise continuous values and the maximum number of values the discrete features take (typically, $B \ll S$). This can be computed in parallel for each feature.

Baseline and Maximum Importance The *permutation importance* of the whole feature set: $e_X(f, \mathbf{s}) = \mathcal{I}_X(f, \mathbf{s}) = \phi_f\left(0, 0, H_X^{\wedge^+}(\mathbf{s}), H_X^{\wedge^-}(\mathbf{s})\right)$ and/or the empty set: $e_\varnothing(f, \mathbf{s}) = \mathcal{I}_\varnothing(f, \mathbf{s}) = \phi_f(0, 0, 0, 0)$ can be added to the info-PI set to improve the model map ϕ_f.

Out-of-distribution Problem In PI, models are evaluated in regions outside the training distribution domain. For CID, substituting PI for permute and retrain or feature ablation solves this issue.

Experimental Section

To test the *CID* ranking adjustment, we first tested it on a toy dataset where the real importances are known, and a real-world medical dataset. We implemented *CID* in Python using scikit-learn's *graphical lasso* (Pedregosa and et al. 2011). For the toy dataset, we used scikit-learn's *Extremely Randomized Trees* and *Bayesian regression* implementations, and for the medical dataset we used a *Gradient Boosting Survival model* (Pölsterl 2020).

Multivariate Generated Data Test

In order to test if *CID* adjusts the permutation ranking into the correct one, we took 2000 samples from a multivariate distribution with the following marginal distributions: $X_1 \sim Uni(0, 1)$, $X_3 \sim$ Gamma(1.5,2), $X_4 \sim$ Beta$(0.5, 0.5)$, $X_2 \sim X_3 \cdot X_4$, $X_5 \sim -$Exponential(0.2), $X_6 \sim \sin(X_4)$ and $X_7 \sim X_8 \cdot X_9 + (1 - X_8) \cdot X_{10}$ with $X_8 \sim$ Bin$(1, 0.7)$ and $X_9 \sim \mathcal{N}(-5, 1)$, $X_{10} \sim \mathcal{N}(5, 1)$. Consider also the binning values: $\mathbf{b} = [0, 0.375, 0.5, 0.575, 0.625, 0.7, 0.775, 0.85, 0.975]$. We then defined the outcome variable as: $y_j = \sum_{i=1}^{7} x_i \cdot \mathbf{1}(b_i \leq u_j < b_{i+1}) + \left(\sum_{k=2}^{4} x_k\right) \cdot \mathbf{1}(b_8 \leq u_j < b_9) + \left(\sum_{l=5}^{6} x_l\right) \cdot \mathbf{1}(u_j \geq b_9)$, where u is an observation of $U \sim$ Uni$(0, 1)$. The true importances are thus: $\mathcal{I}_1 \geq \mathcal{I}_2 \geq \mathcal{I}_3 \geq \mathcal{I}_4 \geq \mathcal{I}_5 = \mathcal{I}_6 \geq \mathcal{I}_7$. We transformed the data into Gaussian using quantile information and chosen gaussian markov random fields to pair with *CID*. The graph was inferred using graphical lasso with a grid-search cross-validation to determine the optimal l_1 penalization parameter. To test the *CID* correction, we performed 200 Shuffle Splits with Extremely Randomized Trees and computed the *Gini importance* for each feature, as well as the *permutation importance*(PI). We then adjusted the feature importances using the *CID* algorithm and Bayesian Regression as ϕ (see assumption 1). You can compare the rankings in figure 3. As can be seen from the swarmplot in figure 3, with the exception of X_1, PI placed a nearly equal weight on all features, centered around zero, presumably due to the high feature covariance. The *CID* was able to rectify this and ranked the features in the right order. It also placed every feature importance at non-zero with a gap between unequally important features and similar importance for

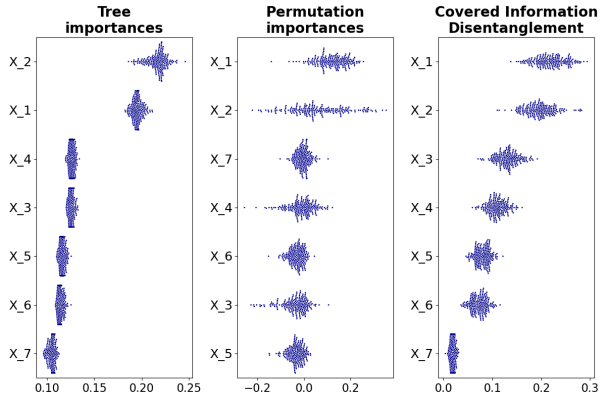

Figure 3: Comparison of the importance ranking on the multivariate gaussian dataset given by from left to right: *Tree importance* (*Gini importance*), *permutation importance*, *CID* importance. The feature order is given by the importance median. The ground truth is $\mathcal{I}_1 \geq \mathcal{I}_2 \geq \mathcal{I}_3 \geq \mathcal{I}_4 \geq \mathcal{I}_5 = \mathcal{I}_6 \geq \mathcal{I}_7$.

X_5/X_6, matching well the true importances. Moreover, notice how the *Gini importance* underestimated X_3/X_1, presumably because X_2 offers many quality splitting points due to the overlap and similarity with X_3/X_4.

Cardiovascular Event Prediction with Proteomics

Problem Introduction Cardiovascular diseases (CVDs) are the number one cause of death globally. Identifying asymptomatic people with the highest cardiovascular (CV) risk remains a crucial challenge in preventing their first cardiac event. Clinically used risk algorithms offer limited accuracy (Piepoli et al. 2016). Consequently, a substantial proportion of the general population at risk remains unidentified until their first clinical event. Hoogeveen and Belo Pereira et al. recently demonstrated increased efficacy in predicting primary events using protein-based models (Hoogeveen et al. 2020). Since technical advances now allow for cheap and reproducible high-throughput proteomic analysis (Assarsson et al. 2014), the field is prime for identifying new diagnostic markers or therapeutic targets, as well as developing new targeted protein panels to quickly and cheaply assess the risk of various diseases. The success of this endeavour is, of course, dependent on reliable feature importance identification.

The reason this dataset is a good candidate to test *CID*, is the "biological robustness" of living systems (Kitano 2004; Stelling et al. 2004). Biological robustness describes a property of living systems whereby specific functions of the system are maintained despite external and internal perturbations. In proteomics, robustness is achieved in two ways: since protein structure is intimately related to function (Schermann 2008), proteins with similar structure can exhibit similar functions, and proteins can be synthesized through different pathways in the metabolic network. This means two proteins located upstream the network relative to a third causing disease will have redundant information, and

Algorithm 1: *CID* Importance

Input: $\mathbf{X}_{M\times N}$, $\mathbf{y}$, f, Ψ, $\mathbf{G}$(optional)
Return: $\mathcal{I}(f)$

$\mathbf{S} \leftarrow$ SampleSubsets($\{1, ..., M\}$)
$\mathbf{e}(f) \leftarrow$ PermutationImportance($\mathbf{X}$, $\mathbf{y}$, f, $\mathbf{S}$)
$\mathbf{G} \leftarrow$ InferGraph($[\mathbf{X}, \mathbf{y}]$) ▷ Infer graph if not povided
$\Psi_\theta \leftarrow$Infer*MRF*Params(Ψ, $\mathbf{X}$, $\mathbf{y}$)
$\mathbf{H}^\wedge \leftarrow$ ComputeMutualInfo($\mathbf{X}$, $\mathbf{y}$), $\mathbf{H^c} \leftarrow \mathbf{0}$
$\mathcal{N} \leftarrow$ GetNeighbors($[\mathbf{X}, y]$, $\mathbf{G}$)
for i **in** $[1, \ldots, N]$ **do** ▷ can be parallelized
  for j **in** $[1, \ldots, M]$ **do**
    $\mathbf{d}, \mathbf{e}, \mathbf{F} \leftarrow$ Potentials(Ψ_θ, $\mathbf{X}$, $\mathbf{y}$, $i, j, \mathcal{N}_i, \mathcal{N}_y$)
    $\mathbf{H}_i^{\mathbf{c}}[j] \leftarrow \mathbf{H}_i^\wedge[j] - log\left(f\frac{\mathbf{d}^T\mathbf{F}\,\mathbf{e}}{\mathbf{d}^T\mathbf{F}_y\mathbf{F}_{x_i}\mathbf{e}}\right)$
  end for
end for
$H^{\mathbf{c}^+}, H^{\mathbf{c}^-}, H^{\wedge^+}, H^{\wedge^-} \leftarrow$ RedundSyn($\mathbf{H}^\wedge, \mathbf{H^c}, \mathbf{S}$)
$\phi \leftarrow$ FitEntropyPI$\left(H^{\mathbf{c}^+}, H^{\mathbf{c}^-}, H^{\wedge^+}, H^{\wedge^-}, \mathbf{e}(f)\right)$
$\mathcal{I}(f) \leftarrow \mathbf{E}_{\sim p(\mathbf{s})}\left[\phi\left(0, H^{\mathbf{c}^-}, H^{\wedge^+}, H^{\wedge^-}\right)\right]$

so do two proteins whose structure is similar (this is depicted in figure 4).

Dataset Description

The dataset consists of a selection of 822 seemingly healthy individuals in a nested case-control sample from the EPIC-Norfolk study (Day et al. 1999). Seemingly healthy individuals were defined as study participants who did not report a history of CV disease. A total of 411 individuals who developed an acute myocardial infarction (either hospitalization or death) between baseline and follow-up through 2016 were selected, together with 411 seemingly healthy individuals who remained free of any CV disease during follow-up. In the original study, the authors demonstrate how predicting short-term events leads to a significant accuracy improvement (Hoogeveen et al. 2020), presumably because the proteomic profile will change over time. We used the early-event prediction dataset, where we only included patients who suffered from an event earlier than 1500 days from measurement (total of 100 patients). We do not make the code for this analysis available due to data confidentiality.

Importance Ranking Experiment Details

To evaluate the models' performance on days-to-event regression, we performed 100 shuffle splits and measured the mean square error on the test set. We used 5-fold cross-validation to select the optimal hyper-parameters of a Survival Gradient Boosting regressor (Pölsterl 2020). To prevent overfitting, we pre-selected 50 proteins using univariate selection. We then compared the *CID* with permutation importance, *Univariate importance* , *SAGE* (Covert, Lundberg, and Lee 2020), and *Tree importance* (*Gini importance*). We used GraphicalLasso (GL) for network inference in all our experiments and selected the l_1 regularization term using grid-search cross-validation. For

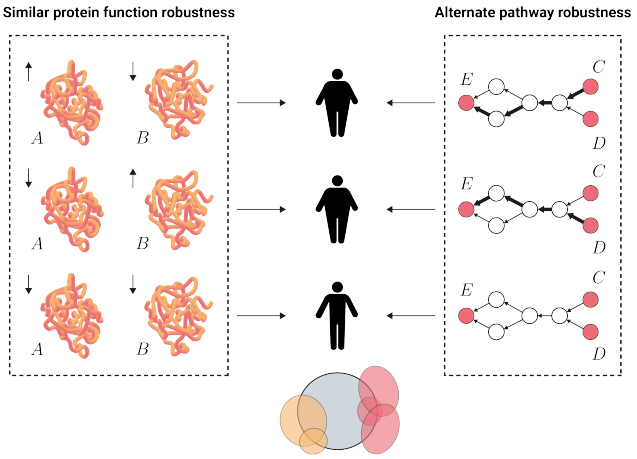

Figure 4: Illustration of biological robustness for the event prediction with proteomics problem. On the left square, it is shown how the levels of two different proteins with similar structure (and hence, similar function) impact the outcome (obesity); on the right square, it is shown how two different proteins can influence the levels of a third outcome-related one through different pathways in the metabolic network; on the bottom, there is a Venn diagram representing the information overlap of the outcome (in gray) and the other proteins considered.

the cardiovascular event survival analysis, we discretized the data into 10 bins. For this experiment we used:

$$\begin{aligned} e_i(f,\mathbf{s}) &= \phi_f\left(H^{\mathbf{c}^+}_{X_i}(\mathbf{s}), H^{\mathbf{c}^-}_{X_i}(\mathbf{s}), H_i^{\wedge^+}(\mathbf{s}), H_i^{\wedge^-}(\mathbf{s})\right) \\ &= \mathcal{I}_i(f,\mathbf{s})g\left(H^{\mathbf{c}^+}_{X_i}(\mathbf{s})\right)\left(1 - \frac{H^{\mathbf{c}^+}_{X_i}(\mathbf{s})}{H_i^{\wedge^+}(\mathbf{s})}\right), \end{aligned}$$

$$g\left(H^{\mathbf{c}^+}_{X_i}(\mathbf{s})\right) = \begin{cases} c, & \text{if } H^{\mathbf{c}^+}_{X_i}(\mathbf{s}) > 0, \quad c \in [1, +\infty[\\ 1, & \text{otherwise} \end{cases},$$

that is, the *permutation importance* is modelled as the true importance weighted by the fraction of uncovered information (disregarding synergy) scaled by c. We then found c using grid-search on the values: $1/c = [1.2, 1.4, 1.6, 1.8, 2, 2.2, 2.4]$. We removed data instances that contained values exceeding 4 times the standard deviation to achieve better discretization.

Results Overall, *CID* spreads the importance more evenly than Perm. imp. and aligns better with the Univariate ranking. Thus, this corroborates the hypothesis that Perm. imp. underrates correlated features. *CID* ranked TRAIL-R2, PSP-D, and IL2-RA two or more places higher, while it ranked SELL and PCOLCE five and seven places lower, respectively.

Gold Standard: To establish a gold-standard analysis of the ranking, we asked world-renowned cardiovascular experts who commented on the comparison. TRAIL-R2 and GDF-15 were identified as the highest predictors of long-term mortality in patients with acute myocardial infarction in

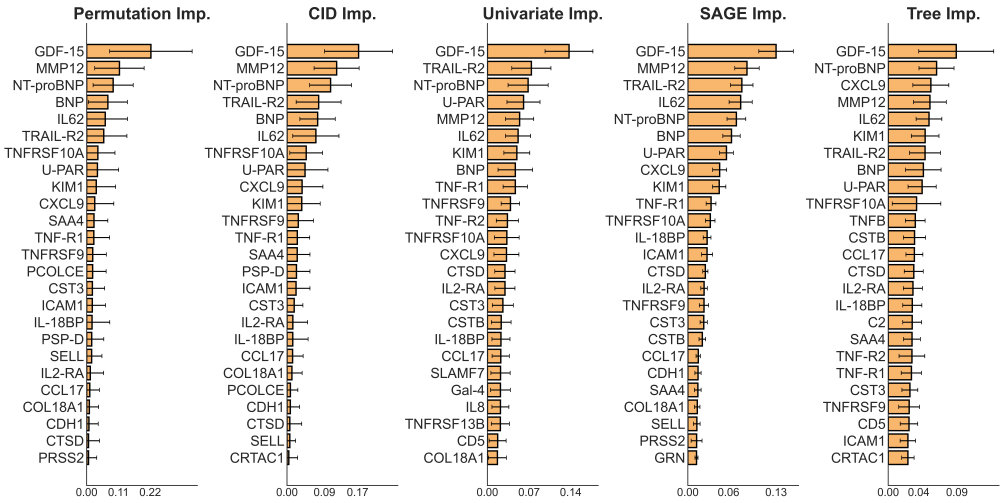

Figure 5: Importance rankings for cardiovascular event prediction using proteomics given by *permutation importance*, *CID*, *univariate importance*, *SAGE* and *tree importance* (*Gini importance*)

Method	Corr.	MSE top feats	Avg. cycle time(s)
Perm. Imp.	0.8697	0.2824 ± 0.0107	0.7756 ± 0.2173
CID	**0.8787**	**0.2801** ± 0.0098	18.76 ± 7.7256
Univar. Imp.	0.8185	0.2947 ± 0.0090	0.0008 ± 0.0003
SAGE	0.8499	0.2858 ± 0.0064	42179 ± 3835
Tree imp.	0.7219	0.2900 ± 0.0087	-

Table 1: Correlation between subset model performance and the subset's sum of importances for each method (higher is better) and the mean squared error on top 10 to 35 features for each method (lower is better), as well as the average running time per cycle in seconds.

(Skau et al. 2017). PSP-D has been identified as a strong clinical predictor of future adverse clinical outcome in stable patients with chronic heart failure in (Brankovic et al. 2019). Il2-RA has been positively associated with all-cause mortality, CVD mortality, incident CVD, stroke, and heart failure in (Durda et al. 2015). To date, SELL and PCOLCE have not been associated as major players in the development of cardiovascular disease.

Quantitative Measure: In order to establish a quantitative measure of the ranking quality, we followed an approach similar to what is described in (Covert, Lundberg, and Lee 2020), where multiple subsets of the data were selected, the models were re-trained for each subset and then for each subset and importance method we measured the correlation between the performance and the subset's sum of importances. We also computed the model performance when trained on the top 10 to 35 proteins of each method. We also report the average running time per cycle conducted on an 8-core Intel(R) Core(TM) i7-7700HQ CPU @ 2.81Ghz. The results are displayed in table 1 which shows *CID* outperformed the other methods on this dataset.

Discussion and Conclusion

Permutation importance is a popular algorithm used to equip black-box models with global explanations. It has the advantage of being easy to understand, but its validity suffers in the presence of covariates. We propose a novel framework (CID) to disentangle the shared information between covariates and show how using *Markov random field*s leads to tractability, making *permutation importance* competitive against methodologies where all marginal contributions of a feature are considered, such as *SHAP*. Due to network inference's complexity, we have only explored *graphical lasso* in conjunction with *Gaussian Markov random field*s. Although this particular implementation is attractive for its scalability and intuitiveness, it might lack sufficient expressive power to model more complex relationships between features.

Recently, A. Fisher proposed *model class reliance* (MCR), a method to estimate the range of variable importance for a pre-specified model class and shown how it can be computed as a series of convex optimization problems for model classes whose empirical loss is convex, although general computation procedures are still an open area of research (Fisher, Rudin, and Dominici 2018). By learning a map between *permutation importance* and entropy terms, the importances retrieved by *CID* are less dependent on the specific fitted model than *permutation importance* or *SHAP*, but the map quality still relies on a consistent model behavior with regards to redundant entropy, as well as a good *MRF* approximation to the data distribution. The former might depend on the groups of features and thus future work includes modeling this map using graph methods on the inferred network, where the node features are the entropy terms. The latter could be improved by using a class of non-parametric *MRF*s with higher flexibility. Should these two problems be solved, then *CID* provides a truly model-agnostic feature importance framework while retaining the intuitiveness of *permutation importance*.

Ethical Impact

With an increasing reliance on using machine learning methods to research impactful domains such as biology and medicine, it is more important than ever to achieve model transparency and accurately determine feature relevance. In this work, we develop an efficient way to incorporate interactions when ranking variables. In the biomedical domain with thousands or millions of complex interactions among proteins, metabolites, genes, and so on, speed and correctness in determining the elements governing a given process are critical because they could significantly mitigate time, resources, and human lives lost. On the other hand, model transparency can also be exploited to develop adversarial examples or gain unwarranted access to protected systems/data.

Acknowledgments

We would like to thank Nick Nurmohamed for helping with the interpretation of the medical dataset results as well as Cláudia Pinhão for the graphics' design.

References

Assarsson, E.; Holmquist, M. L. G.; Björkesten, J.; Thorsen, S. B.; Ekman, D.; Eriksson, A.; Dickens, E. R.; Ohlsson, S.; Edfeldt, G.; Andersson, A.-C.; Lindstedt, P.; Stenvang, J.; Gullberg, M.; and Fredriksson, S. 2014. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PloS one*, 9: e95192.

Brankovic, M.; Akkerhuis, K. M.; Mouthaan, H.; Constantinescu, A.; Caliskan, K.; van Ramshorst, J.; Germans, T.; Umans, V.; and Kardys, I. 2019. Utility of temporal profiles of new cardio-renal and pulmonary candidate biomarkers in chronic heart failure. *International journal of cardiology*, 276: 157–165.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1): 5–32.

Cover, T. M.; and Thomas, J. A. 2012. *Elements of Information Theory*. John Wiley & Sons.

Covert, I.; Lundberg, S.; and Lee, S.-I. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33.

Day, N.; Oakes, S.; Luben, R.; Khaw, K.; Bingham, S.; Welch, A.; and Wareham, N. 1999. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *British journal of cancer*, 80 Suppl 1: 95–103.

Durda, P.; Sabourin, J.; Lange, E. M.; Nalls, M. A.; Mychaleckyj, J. C.; Jenny, N. S.; Li, J.; Walston, J.; Harris, T. B.; Psaty, B. M.; Valdar, W.; Liu, Y.; Cushman, M.; Reiner, A. P.; Tracy, R. P.; and Lange, L. A. 2015. Plasma Levels of Soluble Interleukin-2 Receptor α: Associations With Clinical Cardiovascular Events and Genome-Wide Association Scan. *Arteriosclerosis, thrombosis, and vascular biology*, 35: 2246–2253.

Fisher, A.; Rudin, C.; and Dominici, F. 2018. Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the "Rashomon" Perspective. *Computer Science*.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3).

Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F.; and Giannotti, F. 2018. Local Rule-Based Explanations of Black Box Decision Systems. *http://arxiv.org/abs/1805.10820v1*.

Hoogeveen, R. M.; Pereira, J. P. B.; Nurmohamed, N. S.; Zampoleri, V.; Bom, M. J.; Baragetti, A.; Boekholdt, S. M.; Knaapen, P.; Khaw, K.-T.; Wareham, N. J.; Groen, A. K.; Catapano, A. L.; Koenig, W.; Levin, E.; and Stroes, E. S. G. 2020. Improved cardiovascular risk prediction using targeted plasma proteomics in primary prevention. *European Heart Journal*.

Ince, R. A. 2017. The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv preprint arXiv:1702.01591*.

Kitano, H. 2004. Biological robustness. *Nature Reviews Genetics*, 5(11): 826–37.

Koller, D.; and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Kumar, E.; Venkatasubramanian, S.; Scheidegger, C.; and Friedler, S. A. 2020. Problems with Shapley-value-based explanations as feature importance measures. *Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria*.

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

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774.

Pedregosa, F.; and et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pereira, J.; Groen, A. K.; Stroes, E. S. G.; and Levin, E. 2019. Graph Space Embedding. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI 2019)*, 3253–3259.

Piepoli, M. F.; Hoes, A. W.; Agewall, S.; Albus, C.; andAlberico L Catapano, C. B.; Cooney, M.-T.; Corrà, U.; Cosyns, B.; Deaton, C.; Graham, I.; Hall, M. S.; Hobbs, F. D. R.; Løchen, M.-L.; Löllgen, H.; Marques-Vidal, P.; Perk, J.; Prescott, E.; Redon, J.; Richter, D. J.; Sattar, N.; Smulders, Y.; Tiberi, M.; van der Worp, H. B.; van Dis, I.; Verschuren, W. M. M.; Binno, S.; and Group, E. S. D. 2016. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *European Heart Journal*, 37: 2315–2381.

Pölsterl, S. 2020. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212): 1–6.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).*

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. *AAAI*.

Schermann, J. P. 2008. Amino Acids, Peptides and Proteins. *Spectroscopy and Modeling of Biomolecular Building Blocks*.

Singh, S.; Ribeiro, M. T.; and Guestrin, C. 2016. Programs as Black-Box Explanations. *http://arxiv.org/abs/1611.07579v1*.

Skau, E.; Henriksen, E.; Wagner, P.; Hedberg, P.; Siegbahn, A.; and Leppert, J. 2017. GDF-15 and TRAIL-R2 are powerful predictors of long-term mortality in patients with acute myocardial infarction. *European journal of preventive cardiology*, 24: 1576–1583.

Stelling, J.; Sauer, U.; Szallasi, Z.; III, F. J. D.; and Doyle, J. 2004. Robustness of Cellular Functions. *Cell*, 118(6): 675–85.

Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; and Hothorn, T. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics*, 8: 25.

Štrumbelj, E.; and Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3): 647–665.

Ting, H. K. 2008. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4): 439–447.

Toloşi, L.; and Lengauer, T. 2011. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27: 1986–1994.

Williams, P. L.; and Beer, R. D. 2010. Nonnegative Decomposition of Multivariate Information. *arXiv:1004.2515*.