Marginal Contribution Feature Importance - an Axiomatic Approach for Explaining Data

Amnon Catav [1] **Boyang Fu** [2] **Yazeed Zoabi** [3] **Ahuva Weiss-Meilik** [4] **Noam Shomron** [3] **Jason Ernst** [2 5 6] **Sriram Sankararaman** [2 5 7] **Ran Gilad-Bachrach** [8]

Abstract

In recent years, methods were proposed for assigning feature importance scores to measure the contribution of individual features. While in some cases the goal is to understand a specific model, in many cases the goal is to understand the contribution of certain properties (features) to a real-world phenomenon. Thus, a distinction has been made between feature importance scores that explain a *model* and scores that explain the *data*. When explaining the data, machine learning models are used as proxies in settings where conducting many real-world experiments is expensive or prohibited. While existing feature importance scores show great success in explaining models, we demonstrate their limitations when explaining the data, especially in the presence of correlations between features. Therefore, we develop a set of axioms to capture properties expected from a feature importance score when explaining data and prove that there exists only one score that satisfies all of them, the *Marginal Contribution Feature Importance* (MCI). We analyze the theoretical properties of this score function and demonstrate its merits empirically.

[1]School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel [2]Computer Science Department, University of California, Los Angeles, USA [3]Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel [4]I-Medata AI Center, Tel Aviv Sourasky Medical Center, Tel-Aviv, Israel [5]Department of Computational Medicine, University of California, Los Angeles, USA [6]Department of Biological Chemistry, University of California, Los Angeles, USA [7]Department of Human Genetics, University of California, Los Angeles, USA [8]Department of Biomedical Engineering, Tel-Aviv University, Tel-Aviv, Israel. Correspondence to: Amnon Catav <amnoncatav@mail.tau.ac.il>, Ran Gilad-Bachrach <rgb@tauex.tau.ac.il>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1. Introduction

The increase usage of machine learning has profound influence on many aspects of our lives. Therefore, it is of paramount importance to lessen the black box nature of many machine learning methods (Lipton, 2018). This awareness has led to extensive work on interpretability (Molnar, 2018), explainability (Holzinger et al., 2019), and more specifically regarding feature importance scores (Ribeiro et al., 2016; Lundberg & Lee, 2017; Shrikumar et al., 2017; Sundararajan et al., 2017; Plumb et al., 2018). Many of the previous studies on feature importance focused on assigning importance scores for predictions of a specific trained model. Methods for assigning feature importance scores are often divided into *local* and *global*, where the goal of local scores is to explain how much each feature impacts a specific prediction, while the goal of global scores is to explain how much each feature is impacting the model predictions across the entire data distribution. However, in many scenarios models are used as tools for studying relations in the real world such as the impact of gender on salary or to identify cancer related genes (Jagsi et al., 2012; Danaee et al., 2017; Kothari et al., 2020). In these cases, we care more about the true underlying relations between each feature and the label, rather than the mechanism of a specific trained model. Thus, one can differentiate between feature importance scores that explain the *model* from feature importance scores that explain the *data* (Chen et al., 2020).

To understand the differences between explaining a trained model and explaining the data, consider the case of predicting the existence of a certain health condition using a linear regression model trained over gene expression data. A scientist may be interested in gene importance as a tool for prioritizing the experiments to be done in the lab. Since correlations are common in gene expression data, a phenomenon frequently referred as collinearity is likely to emerge (Mason & Perreault Jr, 1991; Zuber & Strimmer, 2009). In the collinearity setting the coefficients of the model are not uniquely determined and therefore they cannot serve as measures for the significance of features (Dormann et al., 2013). Thus, a scientist trying to understand which genes are most associated with the investigated health condition might get

arbitrary results when using feature importance methods that explain the model. Previous studies have shown that this phenomenon is not restricted only to linear relations. In fact, it may appear whenever the features are correlated (Hooker et al., 2019).

More generally, in the presence of redundant information among features, a model can arbitrary choose to utilize only one or some of the features over the others. However, this arbitrary choice does not necessarily reflect the true relations between these features and the label in the real world. For example, imagine a scientist trying to understand what are the important attributes that contribute to workers' salaries. Assume that these attributes or features include age, education, experience and gender, along with worker's last salary. A common practice today would be to train a model using these features, and to use a feature importance method that describes the model as a measure for the significance of each feature in determining salary. However, it is likely to assume that the last salary feature already encapsulates relevant predictive information from the other features. In this case, the model might arbitrary choose to count almost solely on the last salary feature. Thus, the expectation from a feature importance that explains the model would be to mark only this feature as important. While the scientist is using the model only as a tool for understanding the data, she would expect also the other features to be considered as important if they individually contribute to salary in some sense. Therefore, in this case there is a difference between the expectation from an explanation of the model and the explanation of the true relations between the features and the label in the real world.

Popular methods, such as the ones based on Shapley-Value, are successful as tools for explaining models and their predictions (Shapley, 1953; Lundberg & Lee, 2017; Lundberg et al., 2020; Covert et al., 2020). However, when used for explaining data that contains correlated features, they tend to under-estimate the importance of these features (Kumar et al., 2020). In other words, the addition of features to the system might make similar features to be considered less important. To understand the implications of this phenomenon, consider a biologist suspecting that a certain biochemical pathway may be involved in the development of a certain disease (Alonso et al., 2015). Over sampling metabolites from the suspected pathway might creates a risk that they will be overlooked due to low scores.

As a simple example, consider a system with a single feature that fully describes the label, and a set of K features that are correlated with it. Formally, let $Y = X_1$, and $X_i = 0.5X_1 + \sqrt{0.75}\epsilon_i$ for $i = 2, \dots K$, when $X_1, \epsilon_2, \dots, \epsilon_K \sim \mathcal{N}(0, 1)$ (which yields $\text{Cov}(X_1, X_i) = 0.5$, $\text{VAR}(X_i) = 1$). It can be shown that for all features, including X_1, Shapley-Value[1] will assign a score that is decreasing with K, approaching zero for $K \to \infty$ to all features. Therefore, even X_1 that can predict the outcome perfectly will be marked as having small importance in the presence of many correlated features. In contrast, the MCI method we present in this study is agnostic to K in this setting: it will assigns X_1 an importance score of 1, and a score of 0.25 for all other features. Note that for a large enough K, the average of X_i's for $i > 1$ is similar to X_1 and therefore it is possible to predict Y from these features even in the absence of X_1. Therefore, these features are not considered unimportant.

Unlike previous methods which applied a layer of correction to Shapley-Value, (Chen et al., 2020; Frye et al., 2020b; Wang et al., 2020), we revisit its underlying axioms to propose a new set of axioms that challenges the use of Shapley-Value as a building block. Further, we prove that there is only one function that satisfies these axioms, which we refer to as the *Marginal Contribution Feature Importance* (MCI)[2]. We compare this score to other feature importance scores, both from a theoretical standpoint and from an empirical one, and show that MCI is preferable in explaining the data.

The contributions of this paper are the following: (1) We present three simple properties (axioms) that are necessary for feature importance score in order to explain the data; (2) We prove that these axioms uniquely identify the MCI score; (3) We analyse the theoretical properties of MCI and show empirically that it is more accurate and robust than other available solutions for explaining the data.

2. Problem Formulation

Before continuing the exposition we introduce several notations and conventions. We define a feature as a random variable $f_i : X \mapsto \mathbb{R}$. When it is clear from the context, we use f_i as the value assigned to a point in X. We also denote $F = \{f_1, \dots, f_n\}$ as the set of features used for predicting a certain target label Y. Given the above notations, we define a *feature importance* as a function $I : F \mapsto \mathbb{R}^+$ assigning each feature a score that represents its significance.

"Feature Importance" is an ill-posed term that has many interpretations (Steppe & Bauer Jr, 1997; Breiman, 2001; Ribeiro et al., 2016; Lundberg et al., 2020). In this work we follow Covert et al. (2020), and define the importance of a feature as the amount of *universal predictive power* it contributes. Since this concept is defined in the context of a specific set of features $S \subseteq F$, then the question of evaluating the overall importance of a feature remains.

The universal predictive power of S quantifies the amount of information that can be extracted from S on a target variable

[1] In this example we consider the evalutation function $\nu = 1 - \text{MSE}$ where MSE stands for the Mean Squared Error.

[2] https://github.com/TAU-MLwell/Marginal-Contribution-Feature-Importance

Y. Formally, let $G(S)$ be a set of predictors restricted to use only the features in S. Given a loss function ℓ, the universal predictive power is a function $\nu : \mathcal{P}(F) \mapsto \mathbb{R}^+$, where $\mathcal{P}(F)$ is the power set of F. This function measures the expected loss reduction between the optimal predictor $g \in G(S)$ and the optimal constant predictor $\hat{y} \in G(\emptyset)$:

$$\nu(S) = \min_{\hat{y}} \left(\mathbb{E}[\ell(\hat{y}, Y)]\right) - \min_{g \in G(S)} \left(\mathbb{E}[\ell(g(X), Y)]\right)$$

In importance scores that are designed to explain a model, the class $G(S)$ is defined by projections of the model trained on F to the subset of the features S. However, when trying to explain the data, $G(S)$ can be any class of models that use only the features in S (Covert et al., 2020).

Making the natural assumption that $S_1 \subseteq S_2$ implies that $G(S_1) \subseteq G(S_2)$, the evaluation function ν is guaranteed to be monotonically increasing with respect to S in the sense that $S_1 \subseteq S_2$ implies $\nu(S_1) \leq \nu(S_2)$. This property reflects the intuition that giving more features to the model can only increase the amount of information on the label and thus allows more accurate predictions. We note that it may be challenging to guarantee monotonicity when ν is estimated from data. Therefore, in Theorem 3 we show that in these cases ν is "almost" monotone and this is sufficient to get a good approximation of the feature importance scores.

For simplicity, and to generalize this idea for the terminology used by Shapley-Value, we only require $\nu : \mathcal{P}(F) \mapsto \mathbb{R}^+$ to be monotonically increasing and refer to it as an *evaluation function*. To simplify notation we also assume $\nu(\emptyset) = 0$. Given an evaluation function ν, the task of assigning feature importance scores that are corresponding to the contribution of each feature is not trivial. In the following we review existing feature importance scores using the above notations and discuss their limitations in explaining the data.

3. Previous Studies

We start our discussion with *Shapley-Value* (Shapley, 1953), a fundamental concept in game theory that was recently adopted to the realm of feature importance. Shapley-Value was originally designed for problems of cost allocation where n participants cooperate to achieve a certain good. By treating features as players cooperating to make accurate predictions, this idea was adopted for feature selection (Cohen et al., 2007) and then was extended to local model feature importance by the SHAP method (Lundberg & Lee, 2017), and recently also extended to global model feature importance by the SAGE method (Covert et al., 2020).

Shapley presented four axioms that a fair allocation of cost should have and showed that there is only one fair cost allocation function (Shapley, 1953). According to this function, the importance of each feature f given an evaluation function ν is defined as follows (Covert et al., 2020):

$$I_\nu(f) = \frac{1}{|F|!} \sum_{\sigma \in \pi(F)} \Delta\left(f, S_f^\sigma, \nu\right) \quad . \tag{1}$$

where F is the set of all features, $\pi(F)$ is the set of all permutations of F, S_f^σ is the set of all features preceding f in permutation σ, and $\Delta(f, S, \nu) = \nu(S \cup \{f\}) - \nu(S)$.

Other aspects of using Shapley-Value as a feature importance score are being studied. Many aim at improving the estimation of the Shapley-Value (Aas et al., 2019; Frye et al., 2020a; Williamson & Feng, 2020; Sundararajan & Najmi, 2020). Other studies use Shapley-Value to specifically explain trees (Lundberg et al., 2020), or to derive causal reasoning using additional prior knowledge (Frye et al., 2020b; Wang et al., 2020).

While Shapley-Value axioms make sense in the realm of allocating costs to beneficiaries, several recent works showed how the existence of correlations between the features can harm its adequacy to the data, even when approximated perfectly (Kumar et al., 2020; Frye et al., 2020b). To demonstrate the problem, consider a system with the binary features f_1, f_2, f_3 which are Rademacher random variables such that the target variable is $Y = f_1 \wedge (f_2 \vee f_3)$ and the mutual information is the evaluation function. In this case, Shapley-Value assigns feature importance scores of $0.65, 0.15$, and 0.15 to f_1, f_2, and f_3 respectively. However, if f_1 is duplicated 3 more times, then the feature importance scores become $0.15, 0.18$, and 0.18 for f_1, f_2, and f_3. Note that the importance score of f_1 drops when it is duplicated while the importance score of f_2 and f_3 increases to the point that they become the most important features. This means that if these features were indicators for the presence of a certain protein in a blood sample, then their importance scores may drop when more indicators are measured. As a consequence, if the scientist suspected that a certain mechanism is responsible for a disease, and therefore sampled many proteins that are related to this mechanism, then Shapley-based feature importance scores will suggest that these proteins are of lesser importance.

Ablation studies are commonly used for assigning importance scores to features (Casagrande & Diamond, 1974; Bengtson & Roth, 2008; Hessel et al., 2018). In this method, the importance of a feature is the reduction in performance due to removing this feature. Using the notation above, in Ablation studies the importance of a feature f is $I_\nu(f) = \nu(F) - \nu(F \setminus \{f\})$.

Bivariate association is the complement to Ablation studies. In this method, feature importance is its contribution in isolation, that is, $I_\nu(f) = \nu(\{f\}) - \nu(\emptyset)$. These methods are commonly used in Genome-Wide Association Studies (GWAS) (Liu et al., 2009; Haljas et al., 2018), in feature ranking methods (Zien et al., 2009), feature selection meth-

ods (Guyon & Elisseeff, 2003), or in feature screening methods (Fan & Lv, 2008).

As mentioned in Covert et al. (2020), both Ablation and Bivariate association methods deal imperfectly with specific types of feature interactions. As an example, consider an XOR task where the target variable is the exclusive-or of two Rademacher random variables: $Y = f_1 \oplus f_2$. Bivariate methods would fail to find the association between these features and the target variable. Moreover, if we add an exact duplication of f_1 to the system, an Ablation test would fail to identify f_1 as important, since Y can be predicted perfectly using the other features.

So far, we have discussed importance scores that are model agnostic. However, it is important to mention that there are also importance scores that are specific to a certain type of model. These methods are designed to explain the model, while the goal of this work is to study methods to explain the data. For example, in linear models, it is common to derive importance scores from the coefficient assigned to each feature, while in trees-based models, it is common to look at the sum of the gains from decision nodes (Breiman et al., 1984). In addition, many recent studies focus on explaining neural networks. This includes Integrated Gradients (Sundararajan et al., 2017), DeepLift (Shrikumar et al., 2017), and additional methods (Samek et al., 2017).

In computer vision, feature importance are usually used as a tool to highlight parts of an image on which a model focuses when making a prediction. For example, when making a prediction on the age of a person in an image it is expected that the model will focus on focal points such as the hair or the wrinkles on the side of the eye. In most cases these are feature importances of the local setting since in different images the position of the hair or the wrinkles may be on a different pixel.

We now move forward to introducing Marginal Contribution Feature Importance Method (MCI), which aims to overcome the difficulties of existing methods.

4. Marginal Contribution Feature Importance

In previous sections, we discussed the different scenarios in which feature importance can be used and presented the limitations of current methods in explaining the data. To find a proper score for this scenario, we begin by introducing a small set of properties expected of a feature importance scoring function in this setting. We refer to these properties as axioms. We show that Marginal Contribution Feature Importance (MCI) is the only function that satisfies these axioms, and we study its properties. To introduce these axioms, we define the *Elimination* operation as follows:

Definition 1 *Let F be a set of features and ν be an evaluation function. Eliminating the set $T \subset F$ creates a new set of features $F' = F \setminus T$ and a new evaluation function $\nu' : \mathcal{P}(F)' \mapsto \mathbb{R}^+$ such that $\forall S \subseteq F',\ \ \nu'(S) = \nu(S)$.*

4.1. The Axioms

In the following we introduce a set of axioms that are properties we expect a feature importance score to satisfy:

Definition 2 *A valid feature importance function I_ν for explaining the data is a function $I_\nu : F \mapsto \mathbb{R}^+$ that has the following properties:*

1. ***Marginal contribution**: The importance of a feature is equal or higher than the increase in the evaluation function when adding it to all the other features:*
$$I_\nu(f) \geq (\nu(F) - \nu(F \setminus \{f\})) \quad .$$
2. ***Elimination**: Eliminating features from F can only decrease the importance of each feature.* i.e.*, if $T \subseteq F$ and ν' is the evaluation function which is obtained by eliminating T from F then*
$$\forall f \in F \setminus T, \quad I_\nu(f) \geq I_{\nu'}(f) \quad .$$
3. ***Minimalism**: If I_ν is the feature importance function, then for every function $I : F \mapsto \mathbb{R}^+$ for which axioms 1 and 2 hold, and for every $f \in F$: $I_\nu(f) \leq I(f)$.*

The Marginal contribution axiom requires that if a feature generates an increase of performance even when all other features are present, then its importance is at least as large as the additional gain it creates. This is to say that if the Ablation study (see section 3) shows a certain gain, then the feature importance is at least this gain.

The rationale for the Elimination axiom is that the importance of a feature may be apparent only when some context is present. For example, if the target variable is the XOR of two features, then their significance is apparent only when both are observed. Therefore, eliminating features can cause the estimated importance of the remaining features to drop. On the other hand, if a feature shown to be important, that is, it provides high predictive power given the current set of features, then its predictive power does not decrease when additional features are introduced. Note that it still may be the case that the relative importance of features changes when adding or eliminating features. In other words, the importance score of a feature should not decrease when adding more features to the system. Nevertheless, the size of the increment does not have to be the same for all features and therefore their relative order may change.

Finally, note that if I_ν satisfies the marginal contribution and the elimination axioms, then for every $\lambda > 1$ the function

λI_ν also satisfies these properties. The *Minimalism* axiom provides disambiguation by requiring the selection of the smallest function.

These axioms allow us to present the main theorem which shows the existence and uniqueness of the feature importance function.

Theorem 1 *Let ν be an evaluation function such that ν is non-decreasing. The function*

$$I_\nu(f) = \max_{S \subseteq F} \Delta(f, S, \nu) = \max_{S \subseteq F} (\nu(S \cup \{f\}) - \nu(S)).$$

satisfies the three axioms: (1) marginal contribution, (2) elimination, and (3) minimalism. Furthermore, this function is the only function that satisfies the three axioms.

Theorem 1 shows that there is only one way to define a feature importance function that satisfies the axioms presented above. We call this function the *Marginal Contribution feature Importance* (MCI) score function. Due to space limitations, the proofs of all theorems are provided in the supplementary material in Section A.

4.2. Properties of the Marginal Contribution Feature Importance Function

MCI has many advantageous properties as shown in the following theorem.

Theorem 2 *Let F be a set of features, let ν be an evaluation function and let I_ν be the feature importance function $I_\nu(f) = \max_{S \subseteq F} \Delta(f, S, \nu)$. The following holds:*

- ***Dummy**: if f is a dummy variable, that is $\forall S \subseteq F,\ \Delta(f, S, \nu) = 0$, then $I_\nu(f) = 0$.*
- ***Symmetry**: if f_i and f_j are symmetric, that is if for every $S \subseteq F$ we have that $\nu(S \cup \{f_i\}) = \nu(S \cup \{f_j\})$, then $I_\nu(f_i) = I_\nu(f_j)$.*
- ***Super-efficiency**: $\forall S \subseteq F,\ \nu(S) \leq \sum_{f \in S} I_\nu(f)$.*
- ***Sub-additivity**: if ν and ω are evaluation functions defined on F then $\forall f \in F,\ I_{\nu+\omega}(f) \leq I_\nu(f) + I_\omega(f)$.*
- ***Upper bound the self contribution**: for every feature $f \in F$, $I_\nu(f) \geq \nu(\{f\})$.*
- ***Duplication invariance**: let F be a set of features and ν be an evaluation function. Assume that f_i is a duplication of f_j in the sense that for every $S \subseteq F \setminus \{f_i, f_j\}$ we have that $\nu(S \cup \{f_i, f_j\}) = \nu(S \cup \{f_i\}) = \nu(S \cup \{f_j\})$. If F' and ν' are the results of eliminating f_i then $\forall f \in F',\ I_{\nu'}(f) = I_\nu(f)$.*

Recall that the Shapley-Value is defined by four axioms: efficiency, symmetry, dummy, and additivity (Shapley, 1953). Theorem 2 shows that MCI has the symmetry and dummy properties, but the efficiency property is replaced by a super-efficiency property, while the additivity property is replaced by a sub-additivity property. The upper bound on self contribution shows that MCI always dominates the Bivariate association scores. It is also easy to verify that it upper bounds Shapley-Value and the Ablation scores. Finally, duplication invariance shows that when features are duplicated, feature importance scores do not change. This demonstrates one of the differences between explaining a model and explaining the data. When features are duplicated, a model may use one of the two duplicates, and the explanation should reflect that. However, when explaining the data, since the two features are identical, they must receive equal importance. Note that the same logic may also apply to demonstrate why a feature importance method that explains the data should not be used for feature selection. While the expectation from a feature selection method is to choose only one of the duplicated features, when explaining the data the two features should be considered as equal.

Another interesting property of MCI is the contexts it can provide for the importance of a feature. From the definition of MCI it follows that for every f there is at least one $S \subseteq F$ such that $I_\nu(f) = \Delta(f, S, \nu)$. Every such S is a context with which f provides its biggest gain. In some cases, these contexts can give additional insight to the scientist.

4.3. Computation and Approximation

The complexity of computing the MCI function in a naïve way is exponential in the number of features. Since computing the Shapley-Value is NP-complete (Deng & Papadimitriou, 1994), there is no reason to believe that MCI is easier to compute. In Section C.I of the supplementary material we provide examples for cases where MCI can be computed in polynomial time, for example, when ν is sub-modular. Moreover, like Shapley-Value, MCI can be approximated by sampling techniques (Castro et al., 2009). One interesting property of MCI is that any sampling-based technique provides a lower-bound on the scores. In Section C.II of the supplementary material we also present some upper-bounds that allow saving computations using a branch and bound technique.

Another challenge in computing MCI is obtaining the values of ν for various sets, when only a finite dataset is available. The following theorem shows that the estimates of ν from data uniformly converge as the sample size increases.

Theorem 3 *Let μ be a probability measure over $X \times Y$. Let F be a set of random variables (features) over X. For any $S \subseteq F$ let $\mathcal{H}_S$ be a hypothesis class defined using only*

the features in S and let $d = \log_2 \max_{S \subseteq F}(|\mathcal{H}_S|)$. For any $\epsilon, \delta > 0$ and $m \geq (\frac{2}{\epsilon^2})\left(d + |F| + \log_2\left(\frac{2}{\delta}\right)\right)$ it holds that:

$$P_{D \sim \mu^m}\left[\max_{f \in F} |I_{\nu_D}(f) - I_{\nu}(f)| > \epsilon\right] \leq \delta$$

where the evaluation function $\nu(S)$ is the minimal 0-1 test loss achieved by any hypotheses $h \in \mathcal{H}_S$: $\nu(S) = \min_{h \in \mathcal{H}_\emptyset}\left(E_{(x,y)\sim\mu}(\ell(h(x), y)\right) - \min_{h \in \mathcal{H}_S}\left(E_{(x,y)\sim\mu}(\ell(h(x), y)\right)$ and $\nu_D(S)$ uses the empirical loss on the set D instead of the test-loss.

Simply put, the above theorem states that with high probability, ν can be estimated to within an additive factor using a finite sample and this estimate can be used to approximate MCI to within a similar additive factor. Further details are provided in the supplementary material in Section A.III.

5. Experiments

In this section we analyze the performance of MCI empirically and compare it to other methods. First, we provide two synthetic experiments designed to test the effect that linear and non-linear feature dependencies have on the different methods. Next, two experiments with real world data are presented: one uses gene expression data from the BRCA gene microarray dataset (Tomczak et al., 2015) and the other one uses Electronic Health Records (EHRs) dataset of patients hospitalized with a Bloodstream Infection (BSI) for the task of predicting mortality (Zoabi et al., 2021). In Section E of the supplementary material we present six additional experiments with different datasets from the UCI repository, testing the effect of feature duplication using various model types (Asuncion & Newman, 2007). In the experiments presented in this section MCI is applied with different underlying model types such as Gradient Boosting Trees (Friedman, 2001), Random Forest (Liaw et al., 2002), Support Vector Regression (SVR) (Drucker et al., 1997) and Multi-Layer Perceptron (MLP) (Rumelhart et al., 1985).

To compare MCI to Shapley-Value we follow the proposal of Covert et al. (2020), and apply the Shapley-Value method to the evaluation function $\nu(S)$, where for every S a model is trained using only the features in S. We refer to this method as Shapley-Value. Note that the SAGE method was designed to explain models and therefore it uses the same model to evaluate $\nu(S)$ by marginalizing the effect of the features not in S. For computing the SAGE values we use the public SAGE repository.[3]

5.1. Collinearity Synthetic Experiment

In the following we describe a synthetic experiment designed to compare the different methods when linear correlations exist between the the features. Three I.I.D. random variables are sampled $A, B, C \sim \mathcal{N}(0, 1)$ and define $Y = 3A + 2B + C$. We then create $N_A, N_B, N_C \in \mathbb{N}$ correlated features for A, B and C accordingly. Given N_A, the features $\{f_i^A\}_{i=1}^{N_A}$ are defined such that $f_i^A = A + \mathcal{N}(0, 0.2)$, and the same for f_i^B and f_i^C. We generate two sets of features of the form: $F = \{f_1^A, \ldots, f_{N_A}^A, f_1^B, \ldots, f_{N_B}^B, f_1^C, \ldots, f_{N_C}^C\}$. (1) uncorrelated set F^1, using $(N_A, N_B, N_C) = (1, 1, 1)$; (2) correlated set F^2, using $(N_A, N_B, N_C) = (12, 4, 1)$.

Table 1. **Results of the collinearity experiment.** The Kendall's tau-b correlation coefficient between the ground truth ranking and the scores assigned by the different methods, along with the coefficients of a model trained using the full set of features in each setting (F^1, F^2). Correlation of 1.00 indicates perfect agreement with the ground truth, while -1.00 indicates full disagreement.

Setting	MCI	SV	SAGE	Coefficients
Uncorrelated	**1.00**	**1.00**	**1.00**	**1.00**
Correlated	**0.68**	-0.68	-0.68	-0.68

For $S \subseteq F$ the evaluation function is defined to be $\nu(S) = \mathit{VAR}(Y) - \mathit{MSE}(S)$ where *MSE(S)* is the Mean Squared Error (MSE) over 3-fold cross validation of an Elastic-Net model, trained using only the features in S (Wang et al., 2006). We use regularization coefficient of $\alpha = 0.1$ which yield the lowest MSE among $\alpha \in \{0.01, 0.1, 1.0\}$. We ran MCI, SV and SAGE, and compare them to the ground truth using the Kendall's tau-b correlation coefficient (Kendall & Gibbons, 1990). The Elastic-Net achieves average MSE score of 0.59 and 0.12 when trained using F^1 and F^2 respectively, over 3-fold cross validation. We note that Bivariate obtains the same results as MCI in this experiment while Ablation performs poorly.

As shown in Table 1, all the methods introduce perfect agreement with the ground truth for the uncorrelated set F^1. However, for the correlated set F^2, only MCI handles the data correctly. In fact, as shown in the supplementary material in Section D.I, SV and SAGE completely reverse the expected order and find the feature correlated with C as most important, followed by the features correlated with B, and last by the features correlated with A, despite the fact that $Y = 3A + 2B + C$. In addition, note that the model also reverse its coefficients when trained on F^2 as captured by SAGE.

5.2. Non-Linear Interactions

As a complement to the collinearity experiment presented in Section 5.1, we design an XOR task in which the dependency between the label and the features are not linear. We sample 6 variables $f_1, \ldots, f_6$, from the uniform distribution on $[-1, 1]$, and set a label $Y = \text{sign}(f_1 \times f_2 \times f_3)$. To test the effect of correlations, we create two settings: (1) a

[3] https://github.com/icc2115/sage downloaded 11/2020

set of uncorrelated features: $F^1 = \{f_1 + \varepsilon_1, f_2, \ldots, f_6\}$; (2) a set of features that contains correlations to f_1: $F^2 = F^1 \cup \{f_1 + \varepsilon_2, f_1 + \varepsilon_3\}$. We sample $\varepsilon_1, \varepsilon_2$ and ε_3 independently from $\mathcal{N}(0, 0.05)$.

To estimate $\nu(S)$ for $S \subseteq F$ we use the test accuracy of an MLP consisting of 3 fully connected layers with 8 nodes and tanh activations, followed by a softmax head. Models are trained with a batch size of 512 for 1,000 epochs using early stopping when validation accuracy did not improve for 50 consecutive epochs. The dataset consists of $100K$ examples with a split of $70\%/10\%/20\%$ for train, validation and test. The network achieves test accuracy of 97.2% when trained using F^1 and 98.5% when trained using F^2.

The feature importance scores are presented in Figure 1. The results for the Bivariate method were omitted since it assigned zero score to all features in this experiment. As seen, in the uncorrelated setting the methods are able to identify f_1, f_2 and f_3 as equally important features. However, MCI is the only method that does not diminish the importance of the variables correlated with f_1 in the second setup, when there are 3 features correlated with f_1. It is interesting to note that MCI provides additional insight here when the context for each of the features is observed (context is defined as $argmax_{S \subseteq F} \Delta(f_i, S, \nu)$). The context for all the features correlated with f_1 is $\{f_2, f_3\}$ which reflects the structure of this task.

To test the robustness of the methods for the type of underlying model used, we repeat this experiment with Random Forest. MCI and SV provide similar scores to the ones reported in Figure 1, however, for SAGE the scores does vary due to the change in the underlying model. The details for this experiment are provided in the supplementary material in Section D.II.

5.3. BRCA Experiments

To test the quality of the importance scores in a real world setting, we use a gene analysis task for which scientific knowledge exists about the importance of the features. Specifically, we use the breast cancer sub-type classification task from a gene microarray dataset (BRCA) (Tomczak et al., 2015). This dataset is provided with a set of genes known to be related to breast cancer and therefore expected to be considered as most important. We conduct two experiments: (1) a quality experiment which compares the rankings provided by different methods to the ground truth; (2) an experiment which tests the methods' robustness for changes in the set of genes used in the genomics study.

The BRCA dataset consists of 17,814 genes from 571 patients that have been diagnosed for one of 4 breast cancer sub-types (Tomczak et al., 2015). In each experiment we use a set of 10 genes that were identified as associated with breast cancer as a ground truth (Covert et al., 2020). We denote this set of related genes as R.

Table 2. **Quantitative results for the BRCA experiments.** Upper table presents the NDCG scores for the 1^{st} experimental setup in which quality of scores is measured (higher is better, the perfect score is 1.00). NDCG scores for different top-@k rankings are provided. Bottom table shows results for Experiment II, measuring the robustness of the scores to the addition of random samples of features. The results are the mean $\pm$ SD of the Kendall-tau distance between each pair of rankings when considering only the BRCA related genes. Here lower is better, the perfect score is 0.00.

Experiment I: Quality (NDCG ↑)

Method	@3	@5	@10	@20	@50
MCI	**1.00**	**0.85**	**0.77**	**0.88**	**0.92**
SV	0.77	0.70	0.73	**0.88**	0.88
SAGE	0.77	0.70	0.67	0.73	0.85
Bivariate	**1.00**	**0.85**	**0.77**	0.82	**0.92**
Ablation	0.30	0.21	0.28	0.44	0.61

Experiment II: Robustness

Method	Mean Kendall Distance ↓
MCI	$\mathbf{0.03 \pm 0.03}$
SV	0.23 ± 0.09
SAGE	0.28 ± 0.08
Ablation	0.34 ± 0.11

We define $\nu(S)$ for $S \subseteq F$ to be the average negative log-loss over 3-fold cross validation of a logistic regression model trained using only the features in S. We train each model using the Scikit-learn package, with its default hyper-parameters (Pedregosa et al., 2011). This model achieves average log-loss of 0.77 and average accuracy of 0.74 over 3-folds cross validation, using the genes set defined in the quality experiment.

Since the complexity of computing SV or MCI is exponential, we use the sampling technique proposed by Covert et al. (2020). According to this algorithm, a random set of permutations P_d of the features is sampled. For each $\sigma \in P_d$ we denote by $S_i^\sigma = \{f_j : \sigma(j) < \sigma(i)\}$ and estimate the feature importance for SV and MCI to be:

$$I_\nu^{\text{SV}}(f_i) = \frac{1}{|P_d|} \sum_{\sigma \in P_d} \Delta(f_i, S_i^\sigma, \nu) \text{ and}$$

$$I_\nu^{\text{MCI}}(f_i) = \max_{\sigma \in P_d} \Delta(f_i, S_i^\sigma, \nu)$$

Recall that this method provides an unbiased estimator for the SV and a lower bound for MCI. In our experiments with BRCA, P_d was of the size 2^{15} as we observed that the rankings of both methods stabilizes at this point.

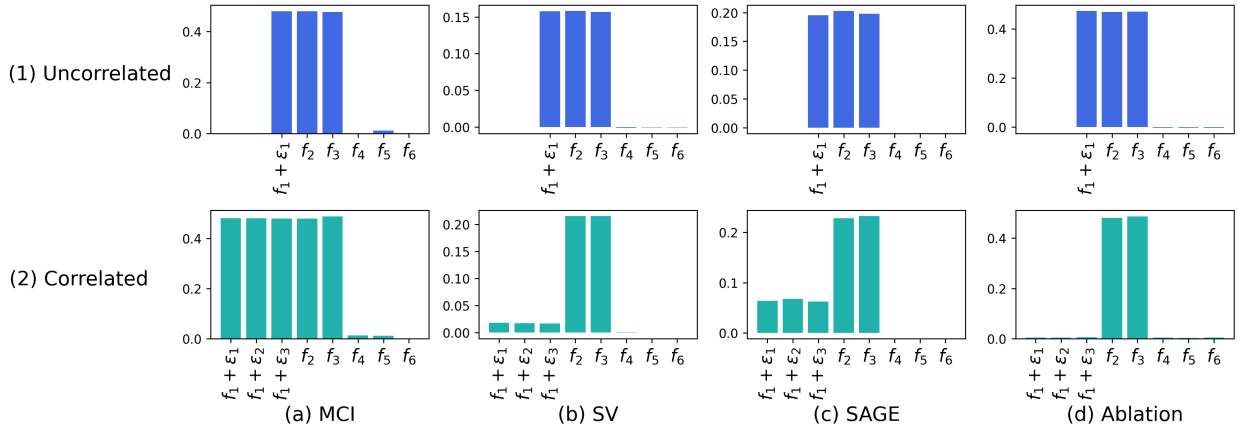

Figure 1. **Results of the non-linear interactions experiment.** The feature importance assigned by the different methods for the 1st and 2nd setups of the experiments are presented in rows 1 and 2. MCI is the only method that does not consider f_1 to be less important when correlated features are added (2nd setup).

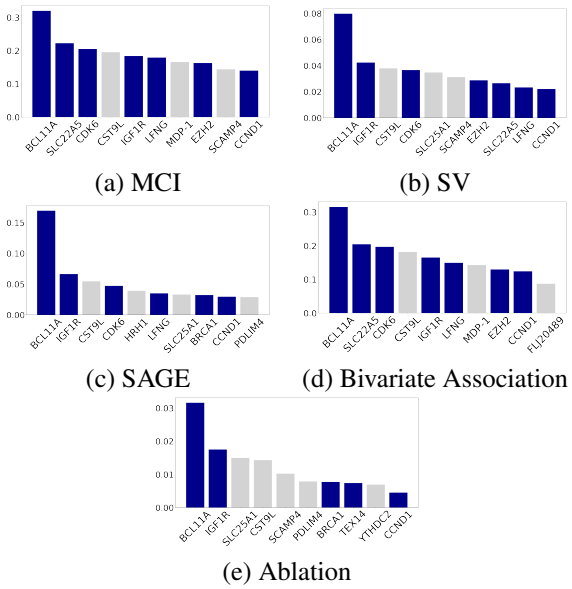

Figure 2. **The top-10 genes ranked by each method in the BRCA quality experiment.** The known BRCA related genes are highlighted in blue. As seen, MCI ranks relevant features higher than most methods.

5.3.1. BRCA Experiment I: Quality

The goal of this experiment is to evaluate the quality of the different methods. For this experiment we use the same subset of 50 genes used by Covert et al. (2020). This subset consists of the set of 10 known BRCA related genes (R) and additional 40 randomly sampled genes. Utilizing the prior scientific knowledge, we consider a score to be better if it ranks the features in R at higher positions. We evaluate the rankings using the Normalized Discounted Cumulative Gain (NDCG) metric (Järvelin & Kekäläinen, 2002).

The results of the experiment are presented in Table 2 and the top-10 ranked genes by each method are shown in Figure 2. The results show that MCI and Bivariate outperform SV and SAGE while Ablation performs poorly. The success of the Bivariate method in this experiment suggests that there are no significant synergies between the features in this dataset which can be detected by logistic regression. MCI handles this situation and even outperforms Bivariate slightly in the top 20 list. However, due to the strong correlations between some of the features, Ablation fails to generate a meaningful ranking and this is a probable explanation also to the low performance of SV and SAGE.

5.3.2. BRCA Experiment II: Robustness

The goal of the second experiment is to evaluate the robustness of the different methods to the list of genes selected to participate in the assessment. To the 10 genes in R we add a sample of additional 40 genes. The 40 genes are sampled at random from the set $G = \{g : \max_{r \in R} |\text{corr}(g, r)| \leq 0.2\}$ ($|G| = 4,596$). We repeat this process 5 times to compute the mean and SD of the Kendall-tau distance (Kendall & Gibbons, 1990) between all pairs of the rankings over R. In this setting, high distances between the rankings suggest high sensitivity to the features sampled, while low distances indicate robustness. In Section D.III of the supplementary material we show that MCI and SV identify all the features in R as the most important for all the samples. This indicates that the randomly added genes have little contribution to the prediction task.

The results of this experiment are presented in Table 2. As seen, MCI produces a consistent ranking for R over the different samples while, SV, SAGE and Ablation produce rankings with higher variance. These results show that the latter methods are sensitive to the inclusion/exclusion of

features, even if they have low correlation with the important features.

5.4. Bloodstream Infection Mortality Experiment

We use another real world dataset to measure the robustness of the different scores (Zoabi et al., 2021). The dataset is extracted from Electronic Health Records (EHRs) of 7,889 patients hospitalized with a Bloodstream Infection (BSI), a condition that directly leads to over 70,000 death cases in the USA annually (Goto & Al-Hasan, 2013). The classification task defined on this data is to predict BSI patients mortality.

Twenty features were extracted from the EHRs by medical experts together with the label that marks mortality. Due to space limitation more details about this dataset are provided in the supplementary materials in Section D.IV.

For this task $\nu(S)$ is defined as the Area Under Curve (AUC) of the Receiver Operator Curve of a gradient boosting trees model trained with LightGBM (version 2.3.0), along with its default parameters (Ke et al., 2017). Since the complexity of SV or MCI is exponential, we use the same sampling technique described in Section 5.3 for 2^{14} random permutations, for which we observed convergence of the rankings.

We compare the importance scores in two settings: (1) the full set of features which empirically contains correlations; (2) a subset of the features, where the features that are highly correlated with Red Blood Cells Count (RBC) are removed. The removed features are Hemoglobin, Hematocrit (HCT), Red Cell Distribution Width (RDW), Mean Cell Hemoglobin (MCH) and Mean Corpuscular Volume (MCV). Note that these features are both statistically correlated (with absolute Pearson correlation in $[0.22, 0.92]$), and also participate in similar biological pathways. The model achieves test AUC score of 0.80 when trained using all the features, and 0.78 without the RBC correlated features.

Both MCI, SV and SAGE rank the RBC feature in the 5-6 position when the correlated features are not present. However, when we add the correlated features, MCI increases the importance of RBC by a factor of 1.32,[4] and promotes its rank by one position, while SV and SAGE decrease the importance of RBC by a factor of 0.54 and 0.53, and demote its rank by one and two positions respectively. Moreover, MCI ranks all of the correlated features both closer and higher (with mean $\pm$ SD rank of 7.3 ± 4.1, 9.8 ± 6 and 10.3 ± 5.8 for MCI, SV and SAGE respectively). We also note that SAGE gives approximately zero importance for Hemoglobin, and a negative score for MCH. This may be explained by the fact that these values can be computed from the other features however, from a clinical point of view these results are troubling.

[4] We apply min-max scaling on each score.

6. Discussion

In this study we investigated feature importance scores as a tool for explaining data (as opposed to explaining a specific model). Such scores may be used by biologists, economists, and sociologist to prioritize their research investments. Since the problem of assigning such scores is ill-posed, we defined a set of properties (axioms) that a feature importance score in this setting is expected to have. We further proved that there is only one score that satisfies all these properties, and we name it the Marginal Contribution feature Importance (MCI). Defining this score using a set of axioms concentrates all the assumptions and expectations from it into a well defined set of properties, and by that allows a way to conduct a healthy debate about its merits and pitfalls. To empirically test this score, we compared MCI to other feature importance scores. We showed that Shapley-Value based methods consider features to be less important if there are other features correlated with them. Therefore, if a scientist samples several features from a pathway she suspects to be important to the studied problem, then Shapley-Value based scores will penalize correlated features, thus contradicting the hypothesis of the scientist. We have conducted an empirical study on synthetic and real-world datasets with different underlying models (GBT, MLP, RF, and logistic regression). The experiments demonstrate that MCI is more robust to changes in the set of features chosen for examination. In the experiments when ground truth was available, we were able to show that MCI is more accurate in ranking higher features that were known to be important. Therefore, we propose the use of MCI as a tool for explaining data.

7. Acknowledgments

We thank Roi Fridman for participating in early stages of this project and Dan Lahav for great discussions. Ran Gilad-Bachrach is supported by a grant from the Tel Aviv University Center for AI and Data Science (TAD) and a grant from the Joy Foundation. Jason Ernst is supported by US National Institutes of Health (DP1DA044371) and a JCCC-BSCRC Ablon Scholars Award. Sriram Sankararaman is supported by NIH grant R35GM125055 and NSF grant III-1705121. Yazeed Zoabi is partially supported by the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. This project was supported by Gilbert Foundation and the TAU-UCLA program for computational biomedical research and training.

References

Aas, K., Jullum, M., and Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *arXiv preprint*

arXiv:1903.10464, 2019.

Alonso, A., Marsal, S., and Julià, A. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology*, 3:23, 2015.

Asuncion, A. and Newman, D. UCI machine learning repository, 2007.

Bengtson, E. and Roth, D. Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 294–303, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D08-1031.

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

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and regression trees*. CRC press, 1984.

Casagrande, V. and Diamond, I. Ablation study of the superior colliculus in the tree shrew (tupaia glis). *Journal of Comparative Neurology*, 156(2):207–237, 1974.

Castro, J., Gómez, D., and Tejada, J. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

Chen, H., Janizek, J. D., Lundberg, S., and Lee, S.-I. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

Cohen, S., Dror, G., and Ruppin, E. Feature selection via coalitional game theory. *Neural Computation*, 19(7): 1939–1961, 2007.

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

Danaee, P., Ghaeini, R., and Hendrix, D. A. A deep learning approach for cancer detection and relevant gene identification. In *Pacific symposium on biocomputing*, pp. 219–229. World Scientific, 2017.

Deng, X. and Papadimitriou, C. H. On the complexity of cooperative solution concepts. *Mathematics of Operations Research*, 19(2):257–266, 1994.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitao, P. J., et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., et al. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

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

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Frye, C., de Mijolla, D., Cowton, L., Stanley, M., and Feige, I. Shapley-based explainability on the data manifold. *arXiv preprint arXiv:2006.01272*, 2020a.

Frye, C., Rowat, C., and Feige, I. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33, 2020b.

Goto, M. and Al-Hasan, M. Overall burden of bloodstream infection and nosocomial bloodstream infection in north america and europe. *Clinical Microbiology and Infection*, 19(6):501–509, 2013.

Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Haljas, K., Amare, A. T., Alizadeh, B. Z., Hsu, Y.-H., Mosley, T., Newman, A., Murabito, J., Tiemeier, H., Tanaka, T., Van Duijn, C., et al. Bivariate genome-wide association study of depressive symptoms with type 2 diabetes and quantitative glycemic traits. *Psychosomatic medicine*, 80(3):242, 2018.

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., and Silver, D. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4): e1312, 2019.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 9734–9745, 2019.

Jagsi, R., Griffith, K. A., Stewart, A., Sambuco, D., DeCastro, R., and Ubel, P. A. Gender differences in the salaries of physician researchers. *Jama*, 307(22):2410–2417, 2012.

Järvelin, K. and Kekäläinen, J. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pp. 3146–3154, 2017.

Kendall, M. and Gibbons, J. D. *Rank Correlation Methods*. A Charles Griffin Title, 5 edition, September 1990.

Kothari, C., Osseni, M. A., Agbo, L., Ouellette, G., Déraspe, M., Laviolette, F., Corbeil, J., Lambert, J.-P., Diorio, C., and Durocher, F. Machine learning analysis identifies genes differentiating triple negative breast cancers. *Scientific reports*, 10(1):1–15, 2020.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pp. 5491–5500. PMLR, 2020.

Liaw, A., Wiener, M., et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

Lipton, Z. C. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

Liu, Y.-Z., Pei, Y.-F., Liu, J.-F., Yang, F., Guo, Y., Zhang, L., Liu, X.-G., Yan, H., Wang, L., Zhang, Y.-P., et al. Powerful bivariate genome-wide association analyses suggest the sox6 gene influencing both obesity and osteoporosis phenotypes in males. *PloS one*, 4(8), 2009.

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

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):2522–5839, 2020.

Mason, C. H. and Perreault Jr, W. D. Collinearity, power, and interpretation of multiple regression analysis. *Journal of marketing research*, 28(3):268–280, 1991.

Molnar, C. Interpretable machine learning. christophm. github. io/interpretable-ml-book. *URL: https://christophm. github. io/interpretable-ml-book*, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Plumb, G., Molitor, D., and Talwalkar, A. S. Model agnostic supervised local explanations. In *Advances in Neural Information Processing Systems*, pp. 2515–2524, 2018.

Ribeiro, M. T., Singh, S., and Guestrin, C. " 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, 2016.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Samek, W., Wiegand, T., and Müller, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

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

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

Steppe, J. M. and Bauer Jr, K. W. Feature saliency measures. *Computers & Mathematics with Applications*, 33(8):109–126, 1997.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *International Conference on Machine Learning*, pp. 9269–9278. PMLR, 2020.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 06–11 Aug 2017.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.

Wang, J., Wiens, J., and Lundberg, S. Shapley flow: A graph-based approach to interpreting model predictions. *arXiv preprint arXiv:2010.14592*, 2020.

Wang, L., Zhu, J., and Zou, H. The doubly regularized support vector machine. *Statistica Sinica*, pp. 589–615, 2006.

Williamson, B. D. and Feng, J. Efficient nonparametric statistical inference on population feature importance using shapley values, 2020.

Zien, A., Krämer, N., Sonnenburg, S., and Rätsch, G. The feature importance ranking measure. *Lecture Notes in Computer Science*, pp. 694–709, 2009. ISSN 1611-3349. doi: 10.1007/978-3-642-04174-7_45. URL http://dx.doi.org/10.1007/978-3-642-04174-7_45.

Zoabi, Y., Kehat, O., Lahav, D., Weiss-Meilik, A., Adler, A., and Shomron, N. Predicting bloodstream infection outcome using machine learning. *medRxiv*, 2021. doi: 10.1101/2021.05.18.21257369. URL https://www.medrxiv.org/content/early/2021/05/19/2021.05.18.21257369.

Zuber, V. and Strimmer, K. Gene ranking and biomarker discovery under correlation. *Bioinformatics*, 25(20):2700–2707, 2009.