From Shapley back to Pearson:
Hypothesis Testing via the Shapley Value

Jacopo Teneggi[*1,3], Beepul Bharti[*2,3], Yaniv Romano[4], and Jeremias Sulam[1,3]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218
[2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218
[3]Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, MD 21218
[4]Department of Electrical Engineering and of Computer Science, Technion, Israel

August, 2022

Abstract

Machine learning models, in particular artificial neural networks, are increasingly used to inform decision making in high-stakes scenarios across a variety of fields—from financial services, to public safety, and healthcare. While neural networks have achieved remarkable performance in many settings, their complex nature raises concerns on their reliability, trustworthiness, and fairness in real-world scenarios. As a result, several a-posteriori explanation methods have been proposed to highlight the features that influence a model's prediction. Notably, the Shapley value—a game theoretic quantity that satisfies several desirable properties—has gained popularity in the machine learning explainability literature. More traditionally, however, feature importance in statistical learning has been formalized by conditional independence, and a standard way to test for it is via Conditional Randomization Tests (CRTs). So far, these two perspectives on interpretability and feature importance have been considered distinct and separate. In this work, we show that Shapley-based explanation methods and conditional independence testing for feature importance are closely related. More precisely, we prove that for binary classification problems, evaluating a Shapley coefficient amounts to performing a specific set of conditional independence tests, as implemented by a procedure similar to the CRT but for a different null hypothesis. Furthermore, the obtained game-theoretic values upper bound the p-values of such tests. As a result, we grant large Shapley coefficients with a precise statistical sense of importance with controlled type I error.

1 Introduction

Deep learning models have shown remarkable success in solving a wide range of problems, from computer vision and natural language processing, to reinforcement learning and scientific research [30]. These exciting results have come hand-in-hand with these predictive models, mostly based on neural networks, becoming extremely complex systems. While these systems consistently set the state-of-the-art in many tasks, our understanding of their mechanics remains intuitive at best. In fact, as neural networks keep getting deeper and wider, they also become *opaque* (or unintelligible) to both developers and users. This lack of transparency is usually referred to as the *black-box* problem: the predictions of a deep learning model are not interpretable. This problem entails both theoretical, societal, and regulatory issues [53, 50, 46]. For example, it is still unclear whether a loss of interpretability is unavoidable to increase performance [41]. Furthermore, it is now being required

---

by law that companies and organizations that rely on autonomous systems provide explanations of the decisions of such systems (e.g. what lead a model to reject a loan application) [26, 19].

These concerns on the reliability, trustworthiness, and fairness of machine learning has motivated recent research efforts on explaining these opaque predictions. A distinction exists between *interpretable* models and *explanation* methods. The former are designed to provide predictions that can be understood by their users. For example, rule-based systems and decision trees are considered interpretable because it is possible to trace a prediction through the specific rules or splits learned by the model. On the other hand, when models are not explicitly designed in an interpretable way (as in deep neural networks), we usually rely on an *a-posteriori* explanation method to find the most important features towards a prediction. These methods assign an importance score to every feature (or group of features) in the input, and the resulting scores can be presented to a user as an *explanation* of the prediction. Explanation methods have been widely adopted given their easy implementation and their ability to explain any model, i.e. they can be *model-agnostic*. Since the introduction of CAM [54], DeepLift [47], and LIME [40], the explainability literature has witnessed remarkable growth, and various approaches have been proposed to identify important features [42, 29, 28, 25, 34, 49, 9, 7] with varying degrees of success [1].

In particular, the SHAP framework [34] unifies previously existing methods while providing explanations that satisfy some desirable theoretical properties. More precisely, it brings the Shapley value [45]—a *solution concept* from game theory [38]—to bear in the interpretation of machine learning predictors. In cooperative game theory, a solution concept is a formal rule that describes the strategy that each player will use when participating in a game. For *Transfer Utility* (TU) games, the Shapley value is the only solution concept that satisfies the properties of additivity, nullity, symmetry, and linearity, and it can be derived axiomatically. In the context of explainability, the SHAP framework considers a TU game represented by the model, where every feature in the input is a player in the game, and it computes its Shapley values as a measure of feature importance. The axiomatic properties of the Shapley value are appealing as they provide interpretable rules that these coefficients satisfy, which may not hold for perturbation-based explanations [44]. An important limitation is that the computational cost of the exact Shapley values is exponential in the number of players (i.e. the input dimension), which quickly makes this intractable for several applications of interest. Thus, most Shapley-based methods rely on various strategies to approximate them [34, 9], some provably [49, 7].

While theoretically sound, and often useful and efficient in practice, this Shapley-based approach to interpretability remains questionable: *What does it mean for a feature to receive a large Shapley value with respect to some predictive model?* Since a precise connection between Shapley values and statistical importance has been missing, it is unclear how one could provide false positive rate control (i.e. controlling Type I error). While the nullity axiom of the Shapley value implies that an unimportant feature will receive a null attribution, it is unclear whether large Shapley values carry some statistical meaning.

In this paper, we answer these questions by relating Shapley values to more traditional definitions of feature importance. Indeed, the statistics literature has a rich history of studying variable selection problems [3, 23, 18], for example, via conditional independence testing and controlled variable selection [6]. That is, a feature is unimportant if a certain response and the feature are independent once conditioned on the remaining features. In other words, knowing the value of an unimportant feature does not provide any more information about the response when the rest of the data is known. The Conditional Randomization Test (CRT) [6] is a conditional independence test that does not make any assumptions on the conditional distribution of the response given the features, however, it assumes that the conditional distribution of the features is known. This setting is particularly useful in applications where unlabeled data may be abundant compared to labeled data (e.g. genomics research, as shown in [6], or imaging data). While the original CRT procedure is computationally intractable for large models, fast and efficient alternatives have been recently proposed, such as the

the Holdout Randomization Test (HRT) [48], and the Distilled Conditional Randomization Test (DCRT) [32].

We remark that in this paper, we define feature importance with respect to the response of a fixed predictive model on an *individual* sample. This differs from the traditional statistical setting in which one analyzes features with respect to an observed response. Nonetheless, this notion of interpretability is important in many scenarios. For example, one may wish to understand the important features for a model that computes credit scores [11, 4], or one may need to verify that the important features for the prediction of an existing, complex model agree with prior-knowledge [5, 25, 13]. We briefly present an overview of our contributions.

1.1 Related work

Some previous works have tried to grant Shapley values with some statistical meaning in certain settings. For example, Owen and Prieur [37], Verdinelli and Wasserman [51] study the Shapley value for correlated variables in the context of ANalysis Of VAriance (ANOVA) [22] and Leave Out COvariates (LOCO) [31] procedures, respectively. Importantly, [51] precisely raises the issue of the lack of statistical meaning for large (or small) Shapley values when features are correlated. Ghorbani et al. [17] build upon [16] and define the *Distributional Shapley* to include the underlying data distribution into the Shapley value. Finally, and most closely to this work, Ma and Tourani [35] also try and deploy ideas of conditional independence to the Shapley value for causal inference on data generated by a Bayesian network. In the current paper, however, we do not make any assumptions on the nature of the distribution of the data, we consider the broader setting of binary classification, and present a novel procedure, inspired by the CRT, that tests for a specific notion of conditional independence of the response of the model.

1.2 Contributions

We will show in this paper that in the context of machine learning models, computing the Shapley value of a given input feature amounts to performing a series of conditional independence tests for specific feature importance definitions. As we will shortly demonstrate, the such computed game-theoretic quantities provide an upper bound to the p-values of the respective conditional independence tests—providing these popular explanation techniques with a well-defined notion of statistical importance. We demonstrate our results with simulated as well as real imaging data, illustrating the tightness of our bound as well as the power of the proposed tests. We do not introduce a novel algorithm to explain model predictions. Instead, we aim to expand our fundamental understanding of what the popular Shapley values compute when applied to machine learning models, and derive new insights that we believe should inform the design of explanation methods contributing to the responsible use of machine learning models.

2 Background

Before presenting the contribution of this work, we briefly introduce the necessary notation and general definitions. Herein, we will denote random variables with capital letters (e.g. X), their realizations with lowercase letters (e.g. $X = x$), and random matrices with bold capital letters (e.g. $\mathbf{X}$). Let $\mathcal{X}$, $\mathcal{Y}$ be some input and response domains, respectively, such that $(X, Y) \sim \mathcal{D}$ is a labeled sample from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We set ourselves within the classical supervised learning framework for binary classification. Let $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{Y} = \{0, 1\}$, and the task is to estimate the binary response Y on a sample X. The distribution $\mathcal{D}$ is usually unknown, and one instead has access to a set $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^m$ of m i.i.d. samples drawn from $\mathcal{D}^m$, with which one can choose a predictor $\hat{f}$ that minimizes the empirical risk over S. In this work, we will not concern ourselves with this optimization problem. Instead, we will assume we are given a predictor, $\hat{f}$, which we wish to analyze. More precisely, for a given input x, we want to understand the importance of the features in x with respect to the value of $\hat{f}(x)$. This question will be formalized by means of the Shapley value, which we now define.

2.1 The Shapley value

In game theory [38], the tuple $([n], v)$ represents an n-person cooperative game with players $[n] :=$ $\{1, \ldots, n\}$ and characteristic function $v : \mathcal{P}([n]) \to \mathbb{R}^+$, where $\mathcal{P}(n)$ is the power set of $[n]$. The characteristic function, $v(C)$, is a nonnegative score that represents the utility accumulated by the players in C, $\forall C \subseteq [n]$. Furthermore, one typically assumes $v(\emptyset) = 0$. A transfer utility (TU) game is one where players can exchange their utility without incurring in any cost. Given a TU game $([n], v)$, a *solution concept* is a formal rule that assigns to every player in the game a reward that is commensurate with their contribution. In particular, there exists a unique solution concept that satisfies the axioms of additivity, nullity, symmetry, and linearity [45] (see Appendix B for details on the axioms). This solution concept is the set of Shapley values, $\phi_1([n], v), \ldots, \phi_n([n], v)$, which are defined as follows.

Definition 1 (Shapley value). Given an n-person TU game $([n], v)$, the Shapley value is

$$\phi_j([n], v) = \sum_{C \subseteq [n] \setminus \{j\}} \frac{|C|! \cdot (n - |C| - 1)!}{n!} \cdot [v(C \cup \{j\}) - v(C)], \ \forall j \in [n]. \tag{1}$$

That is, $\phi_j([n], v)$ is the average marginalized contribution of the jth player over all coalitions of players. We remark that the number of coalitions of players grows exponentially with n.

2.2 Explaining model predictions with the Shapley value

As defined in Definition 1, these coefficients are computed for players in a TU game, and it is not immediately clear how this would apply to machine learning models. To this end, let $\hat{f}(x)$ be the prediction of a learned model on a new sample $x \sim \mathcal{D}_\mathcal{X}$. For any set of features $C \subseteq [n]$, denote $x_C \in \mathbb{R}^{|C|}$ the entries of x in the subset C (analogously, $x_{-C} \in \mathbb{R}^{n-|C|}$ is its complement), and $M_C(x) = [x_C, \widetilde{X}_{-C}] \in \mathbb{R}^n$ the random vector that is equal to x in the features in C and that takes an *unimportant reference* value in its complement. Following previous literature [34], we let $\widetilde{X}_{-C}$ be sampled from its conditional distribution given X_C, i.e. $\widetilde{X}_{-C} \sim X_{-C} \mid X_C = x_C$.

Given a predictor $\hat{f}$, define v to be the characteristic function such that $v(C) = \hat{g}(M_C(x)) = \mathbb{E}\left[\hat{f}(M_C(x))\right]$. This way, one can define the TU game $(x, \hat{g})$[1], where every feature in the sample x is a player in the game. Then, following Definition 1, the Shapley value of $(x, \hat{g})$ becomes

$$\phi_j(x, \hat{g}) = \sum_{C \subseteq [n] \setminus \{j\}} \frac{|C|! \cdot (n - |C| - 1)!}{n!} \cdot \left[\hat{g}(M_{C \cup \{j\}}(x)) - \hat{g}(M_C(x))\right], \ \forall j \in [n]. \tag{2}$$

In the context of explainability, for example, we look for a subset of features $C^* \subseteq [n]$ such that x_{C^*} are the observed features that contributed the most towards $\hat{f}(x)$. Then, C^* contains the features with a positive (or high enough) Shapley coefficient, i.e. $C^* = \{j \in [n] : \phi_j(x, \hat{g}) > 0\}$.

We remark that a central difference between the original definition of the Shapley value from game theory (Eq. (1)) and its translation to machine learning models (Eq. (2)) is that the model $\hat{f}$ has a fixed input domain $\mathcal{X} \subset \mathbb{R}^n$, and furthermore it was trained on samples from a specific distribution over that domain, i.e. the marginal distribution $\mathcal{D}_\mathcal{X}$. Thus, one must mask features with a reference value $\widetilde{X}_{-C}$ from a specific conditional distribution so that $M_C(x) \sim \mathcal{D}_\mathcal{X}$. Lastly, both the need to approximate the conditional distribution with high accuracy and the exponential number of model evaluations required to compute the exact Shapley values make them intractable in most scenarios of interest. In certain cases, provable approximations exist [49, 7].

2.3 The Conditional Randomization Test (CRT)

We change gears to define variable importance from the perspective of conditional independence. Let X, Y, and Z be random variables, we say that Y is independent of X given Z (succinctly,

---

[1]With abuse of notation for simplicity.

$Y \perp\!\!\!\perp X \mid Z)$ if

$$(Y \mid X = x, \ Z = z) \overset{d}{=} (Y \mid Z = z), \tag{3}$$

where $\overset{d}{=}$ indicates equality in distribution. Recall that $\forall j \in [n]$, $X_j \in \mathbb{R}$ is a random variable that corresponds to the jth feature in X, and $X_{-j} \in \mathbb{R}^{n-1}$ is its complement. Then, the CRT [6] implements a conditional independence test for the null hypothesis

$$H_{0,j}^{\text{CRT}} : \ Y \perp\!\!\!\perp X_j \mid X_{-j}. \tag{4}$$

Note that from the definition of conditional independence, the null hypothesis $H_{0,j}^{\text{CRT}}$ states that

$$(X_j \mid Y = y, X_{-j} = x_{-j}) \overset{d}{=} (X_j \mid X_{-j} = x_{-j}). \tag{5}$$

Denote $\widetilde{X}_j^{(i)} \sim X_{-j} \mid X_j$, $i = 1, \ldots, K$, null[2] duplicates of X_j. Under $H_{0,j}^{\text{CRT}}$, for any choice of test statistic $T(X_j, Y, X_{-j})$, the random variables $T(X_j, Y, X_{-j}), T(X_j, Y, \widetilde{X}_{-j}^{(i)}), \ldots, T(X_j, Y, \widetilde{X}_{-j}^{(K)})$ are i.i.d, hence exchangeable [2]. Then, the p-value returned by the CRT is valid: under $H_{0,j}^{\text{CRT}}$, $\mathbb{P}[\hat{p}_j^{\text{CRT}} \leq \alpha] \leq \alpha$, $\forall \alpha \in [0,1]$ [see 6, Lemma F.1]. For completeness, Algorithm C.1 summarizes the CRT procedure for a random data matrix $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(m)}) \in \mathbb{R}^{m \times n}$, $\boldsymbol{Y} = (Y^{(1)}, \ldots, Y^{(m)}) \in \{0,1\}^m$ of m i.i.d. random variables sampled from $\mathcal{D}^m$.

3 Hypothesis testing via the Shapley value

Given the above background, we will now see that while seemingly unrelated, Shapley values are tightly connected to conditional independence tests. Formally, we present a novel modification of the CRT that, differently from the original motivation in Candes et al. [6], evaluates conditional independence of the response of a deterministic model $\hat{f}$ on a sample x.

3.1 The SHAPley Local Independence Test (SHAPLIT)

We now introduce the SHAPley Local Independence Test (SHAPLIT), a new variation of the CRT that tests for the independence of the output of the model $\hat{f}$ on x_j, when the features in x_C, $C \subseteq [n] \setminus \{j\}$ are present. We now make this procedure precise, employing the notation defined thus far.

Definition 2 (SHAPley Local Independence Test (SHAPLIT)). Let $\mathcal{X} \subseteq \mathbb{R}^n$, $\hat{f} : \mathcal{X} \to [0,1]$ be a model trained on data $S \sim \mathcal{D}^m$. Given a new sample $x \sim \mathcal{D}_{\mathcal{X}}$, $\forall C \subseteq [n]$, denote with $M_C(x)$ the corresponding masked random vector. Given a feature $j \in [n]$, for any set $C \subseteq [n] \setminus \{j\}$, the null hypothesis is

$$H_{0,j,C}^{\text{SHAPLIT}} : \ \hat{f}\left(M_{C \cup \{j\}}(x)\right) \overset{d}{=} \hat{f}\left(M_C(x)\right), \tag{6}$$

where $\overset{d}{=}$ denotes equality in distribution.[3]

Similarly to the HRT [48] and the IRT [5], our procedure does not require to train several predictors, and it is defined *locally* on a sample x rather than over a population. Furthermore, the SHAPLIT procedure differs from the IRT [5] in that it tests for the conditional independence of the response of the model with respect to a feature $j \in [n]$ rather than a group of subsets of features, and it allows to condition on arbitrary subsets $C \subseteq [n] \setminus \{j\}$ instead of only considering $-j$, i.e. all features but the jth one. Recall from Section 2.2 that $M_C(x) = [x_C, \widetilde{X}_{-C}] \in \mathbb{R}^n$ is a random vector

---

[2]We denote $\widetilde{X}_j^{(i)}$ as *null* because, by construction, they are conditionally independent of Y given X_j. In other words, they come from the null distribution.

[3]Note that the SHAPLIT procedure can be generalized to models define on arbitrary domains $\mathcal{X}$ and $\mathcal{Y}$. In the context of this work, we consider binary classification.

Algorithm 1 Shapley Local Independence Test

---

    procedure SHAPLIT(model $\hat{f} : \mathbb{R}^n \to [0, 1]$, sample $x \in \mathbb{R}^n$, feature $j \in [n]$, subset $C \subseteq [n] \backslash \{j\}$, test statistic T, number of null draws $K \in \mathbb{N}$, number of reference samples $L \in \mathbb{N}$)

        Sample $\widetilde{X}_{-(C \cup \{j\})}^{(i)} \sim X_{-(C \cup \{j\})} \mid (X_C = x_C, X_j = x_j),\ i = 1, \ldots, L$

        $\widehat{\boldsymbol{Y}}_{C \cup \{j\}} \leftarrow (\hat{f}([x_{C \cup \{j\}}, \widetilde{X}_{-(C \cup \{j\})}^{(1)}]), \ldots, \hat{f}([x_{C \cup \{j\}}, \widetilde{X}_{-(C \cup \{j\})}^{(L)}]))$

        Compute the test statistic, $t \leftarrow T(\widehat{\boldsymbol{Y}})$

        for $k \leftarrow 1, \ldots, K$ **do**

            Sample $\widetilde{X}_{-C}^{(i)} \sim X_{-C} \mid X_C = x_C,\ i = 1, \ldots, L$

            $\widehat{\boldsymbol{Y}}_C^{(k)} \leftarrow (\hat{f}([x_C, \widetilde{X}_{-C}^{(1)}]), \ldots, \hat{f}([x_C, \widetilde{X}_{-C}^{(L)}]))$

            Compute the null statistic, $\tilde{t}^{(k)} \leftarrow T(\widehat{\boldsymbol{Y}}_C^{(k)})$

        end for

        return A (one-sided) p-value

$$\hat{p}_{j,C}^{\mathrm{SHAPLIT}} = \frac{1}{K+1} \left(1 + \sum_{k=1}^{K} \mathbf{1} \left[\tilde{t}^{(k)} \geq t \right] \right)$$

    end procedure

---

equal to x in the features in C, and that takes an unimportant (random) reference value—sampled from its conditional distribution—in its complement. This way, one can rephrase the null hypothesis in Eq. (6) as

$$\left(\hat{f}([x_{C \cup \{j\}}, \widetilde{X}_{-(C \cup \{j\})}]) \mid X_{C \cup \{j\}} = x_{C \cup \{j\}} \right) \overset{d}{=} \left(\hat{f}([x_C, \widetilde{X}_{-C}]) \mid X_C = x_C \right). \tag{7}$$

How should one test for this null? We now introduce a novel procedure inspired by the CRT that precisely carries out this test. More precisely, $\forall C \subseteq [n]$, denote $\boldsymbol{M}_C(x) = (M_C^{(1)}(x), \ldots, M_C^{(L)}(x)) \in \mathbb{R}^{L \times n}$ the random matrix containing L i.i.d. samples from $M_C(x)$, such that $\widehat{\boldsymbol{Y}}_C = \hat{f}(\boldsymbol{M}_C(x)) \in \mathbb{R}^L$ are the predictions of $\hat{f}$ on each $M_C^{(i)}(x)$. For any choice of test statistic $T(\widehat{\boldsymbol{Y}}_C)$, e.g. the mean, Algorithm 1 implements the SHAPLIT procedure. We remark that although the SHAPLIT procedure is valid for $L = 1$, one may want to sample $L > 1$ random masked vectors to account for the risk of the model. We now formally state the validity of this test.

Theorem 1 (Validity of $\hat{p}_{j,C}^{\mathrm{SHAPLIT}}$). *Under the null hypothesis $H_{0,j,C}^{SHAPLIT}$ in Eq. (6), the p-value returned by the SHAPLIT procedure is valid for any choice of test statistic T, i.e. $\mathbb{P}[\hat{p}_{j,C}^{SHAPLIT} \leq \alpha] \leq \alpha,\ \forall \alpha \in [0, 1]$.*

    See Appendix A.1 for the short proof of this result, and let us make a remark about the involved distributions. Recall that the SHAPLIT procedure is defined locally on a sample x. Intuitively, under the null $H_{0,j,C}^{\mathrm{SHAPLIT}}$, the distribution of the response of the model $\hat{f}$ is independent of x_j (the *observed* value of the jth feature of x) conditionally on $X_C = x_C$. Since $\hat{f}$ is a deterministic function of its input, it is natural to ask when the random variable $\hat{f}(M_{C \cup \{j\}}(x))$ has a degenerate distribution. Assume first that $\forall C \subset [n],\ X_{-C} \mid X_C = x_C$ is not degenerate (e.g. it is not constant), otherwise $\hat{f}(M_{C \cup \{j\}}(x))$ is trivially degenerate. Then, when $C = -j$ (that is, all features but the jth one) the complement set is empty, i.e. $-(C \cup \{j\}) = \emptyset$, and $\hat{f}(M_{C \cup \{j\}}(x)) \equiv \hat{f}(x)$. That is, the distribution of $\hat{f}(M_{C \cup \{j\}}(x))$ is point mass at $\hat{f}(x)$. In this case, sampling only one random masked vector (i.e. $L = 1$ in the SHAPLIT procedure) retrieves the IRT procedure of Burns et al. [5] when testing only for the subset $\{j\}$.

3.2 Connecting the Shapley value and the SHAPLIT conditional independence test

The attentive reader will have noticed that the Shapley value comprises differences of means of pairs of distributions which are related to the nulls of SHAPLIT. We now make this connection more precise, which will allow us to associate the marginal contributions in the Shapley value with the p-value returned by SHAPLIT.

Let $L = 1$ in the SHAPLIT procedure and let $T(\cdot)$ be the identity, such that given a new sample $x \sim \mathcal{D}_{\mathcal{X}}$, and a fixed feature $j \in [n]$, $\forall C \subseteq [n] \setminus \{j\}$, the test statistic is $t = \hat{f}(M_{C \cup \{j\}}(x))$, and the null statistic is $\tilde{t} = \hat{f}(M_C(x))$. Assume that $\hat{f}(x)$ is large for important features (otherwise, consider $1 - \hat{f}(x)$), and define the random variable

$$\Gamma_{j,C} := \hat{f}(M_{C \cup \{j\}}(x)) - \hat{f}(M_C(x)) \tag{8}$$

such that $\gamma_{j,C} = \mathbb{E}[\Gamma_{j,C}] = \mathbb{E}[\hat{f}(M_{C \cup \{j\}}(x))] - \mathbb{E}[\hat{f}(M_C(x))]$ and $\lim_{K \to \infty} \hat{p}_{j,C}^{\text{SHAPLIT}} = \mathbb{P}[\Gamma_{j,C} \leq 0]$. Furthermore, $\forall C \subseteq [n]$, denote $\hat{g}(M_C(x)) = \mathbb{E}[\hat{f}(M_C(x))]$ and recall that the tuple $(x, \hat{g})$ defines an n-person TU game such that we can rewrite the definition of its Shapley value in Eq. (2) as

$$\phi_j(x, \hat{g}) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \left(\hat{g}(M_{C \cup \{j\}}(x)) - \hat{g}(M_C(x)) \right) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \gamma_{j,C}, \tag{9}$$

for some appropriate constants w_C. We now show that each summand $\gamma_{j,C}$ in $\phi_j(x, \hat{g})$ upper bounds the p-value of the respective SHAPLIT test, $p_{j,C}$.

Theorem 2. *Let $\mathcal{X} \subseteq \mathbb{R}^n$, $\hat{f} : \mathcal{X} \to [0,1]$ be a trained model on data $S \sim \mathcal{D}^m$. Given a sample $x \sim \mathcal{D}_{\mathcal{X}}$ and a feature $j \in [n]$, $\forall C \subseteq [n] \setminus \{j\}$, define $\Gamma_{j,C} := \hat{f}(M_{C \cup \{j\}}(x)) - \hat{f}(M_C(x))$ such that $\gamma_{j,C} = \mathbb{E}[\Gamma_{j,C}]$ and $p_{j,C} = \mathbb{P}[\Gamma_{j,C} \leq 0]$. Then,*

$$p_{j,C} \leq 1 - \gamma_{j,C}. \tag{10}$$

Note that while the bound above is on expectations, an analogous result can be easily derived by using finite-sample bounds and approximating $\gamma_{j,C}$ with an empirical mean (as long as these are computed with samples independent of $\hat{f}$). We defer the proof of this result to Appendix A.2, and provide now a few remarks.

First, note that Theorem 2 provides a novel understanding of Shapley coefficients from a conditional independence testing perspective. In particular, each summand $\gamma_{j,C}$ in $\phi_{j,C}(x, \hat{g})$ is an upper bound on the p-value of the conditional independence test with null hypothesis $H_{0,j,C}^{\text{SHAPLIT}}$, i.e. $\hat{f}(M_{C \cup \{j\}}(x)) \overset{d}{=} \hat{f}(M_C(x))$.

Second, it is easy to see that if $\hat{f}(x) = 1$, then $\phi_j(x, \hat{g}) = 1 \iff \gamma_{j,C} = 1$, $\forall C \subseteq [n] \setminus \{j\}$. Until now, it was unclear whether large Shapley values carried any statistical meaning. Theorem 2 precisely shows that a Shapley value of 1 can be only attained if all 2^{n-1} tests in $\phi_j(x, \hat{g})$ reject their nulls. We now extend this observation by providing an upper bound on the p-value of all tests in the Shapley value when $\phi_j(x, \hat{g}) \geq 1 - \epsilon$, $\epsilon \in (0, 1)$.

Corollary 1. *In the setting of Theorem 2, assume that for a feature $j \in [n]$, $\phi_j(x, \hat{g}) \geq 1 - \epsilon$, $\epsilon \in (0, 1)$. Let $\widetilde{w} = \min_{C \subseteq [n] \setminus \{j\}} w_C$. Then, $\forall C \subseteq [n] \setminus \{j\}$,*

$$\gamma_{j,C} \geq \frac{\widetilde{w} - \epsilon}{\widetilde{w}}. \tag{11}$$

Which implies that $\forall C \subseteq [n] \setminus \{j\}$, $p_{j,C} \leq 1 - (\widetilde{w} - \epsilon)/\widetilde{w}$.
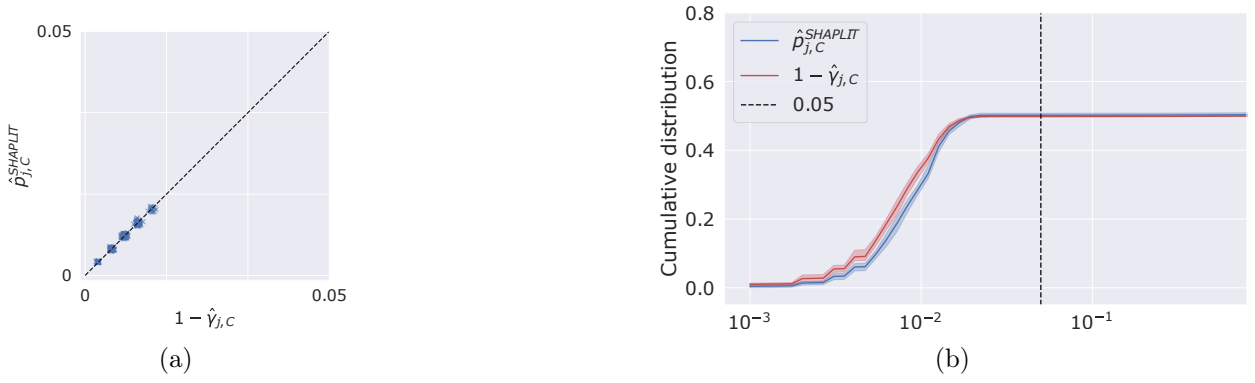
Figure 1: (a) Estimates of $p_{j,C}^{\text{SHAPLIT}}$ and $\gamma_{j,C}$ for all 512 tests in the Shapley value of one important feature j. (b) The average empirical cumulative distributions of $\hat{p}_{j,C}^{\text{SHAPLIT}}$ and $1 - \hat{\gamma}_{j,C}$ on $K = 1000$ samples for all 512 tests for 20 important features. The black dotted vertical line marks the critical value of 0.05.

This result, whose proof we include in Appendix A.3, shows that a large enough $\phi_j(x, \hat{g})$ implies that the p-values of *all* SHAPLIT tests cannot be larger than a certain (small) amount, granting $\phi_j(x, \hat{g})$ with statistical meaning. Note, however, that since there exist 2^{n-1} summands in $\gamma_{j,C}$, ϵ must decrease exponentially fast with n for the bound above to be informative.[4] One can also naturally include a correction for multiple hypothesis testing, such as the Bonferroni method [43], to account for Type I error inflation due to multiple hypotheses testing [5].

Third, and conversely to the points above, note that these results also provide a new interpretation to a *low* Shapley summand, $\gamma_{j,C}$. While in the context of game theory these players "do not contribute much to the outcome of the game", the bound in Theorem 2 simply implies that one cannot reject the null, as the upper bound on the p-values will be rather large. Finally—and related to this point—we note that features with negative contributions, i.e. $\gamma_{j,C} \leq 0$, make the bound in Theorem 2 vacuous since $p_{j,C} \leq 1 \leq 1 - \gamma_{j,C}$ always. However, if one is interested in finding such features, it suffices to use $t = 1 - \hat{f}(M_{C \cup \{j\}}(x))$ and $\tilde{t} = 1 - \hat{f}(M_C(x))$ such that $p_{j,C} = \mathbb{P}[\Gamma_{j,C} \geq 0]$, and the analogous version of Theorem 2 follows.

3.3 The Shapley value as a multiple hypothesis test

Given k null hypotheses $H_{0,1}, \ldots, H_{0,k}$, an important task in multiple hypothesis testing is *global* testing, where one tests the *global* null hypothesis $H_0 = \bigcap_{i=1}^k H_{0,i}$. The global null is true if and only if every individual null hypothesis is true. There exists extensive literature in the field of mathematical statistics on combining the k p-values $p_{0,1}, \ldots, p_{0,k}$ of the individual tests in some valid fashion to test the global null [21, 15, 52]. In our setting, for a feature $j \in [n]$, there exists a SHAPLIT test with null hypothesis $H_{0,j,C}^{\text{SHAPLIT}}$ for every $C \subseteq [n] \setminus \{j\}$. Therefore, we can define our global null hypothesis to be

$$H_{0,j}^{\text{global}} = \bigcap_{C \subseteq [n] \setminus \{j\}} H_{0,j,C}^{\text{SHAPLIT}}. \tag{12}$$

Indeed, we now introduce the following corollary of Theorem 2 showing that the Shapley value of feature j can be used to test for $H_{0,j}^{\text{global}}$.

Corollary 2. *Denote $p_j^{global} = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot p_{j,C}$ the weighted arithmetic mean of the p-values of all SHAPLIT tests for feature $j \in [n]$ and let ϕ_j be its Shapley value. Then, under $H_{0,j}^{global}$, $2 \cdot p_j^{global}$ is a valid p-value, and*

$$p_j^{global} \leq 1 - \phi_j. \tag{13}$$

The proof is included in Appendix A.4. This result implies that, given a desired level α, we can use ϕ_j to reject $H_{0,j}^{\text{global}}$ when $\phi_j \geq 1 - \alpha/2$. In light of Corollary 1, this condition on ϕ_j is satisfied

---

[4]In particular, $p_{j,C} \leq \alpha'$, $\alpha' = \alpha/2^{n-1}$ implies that $\epsilon = O(\alpha/\sqrt{n}4^n)$.
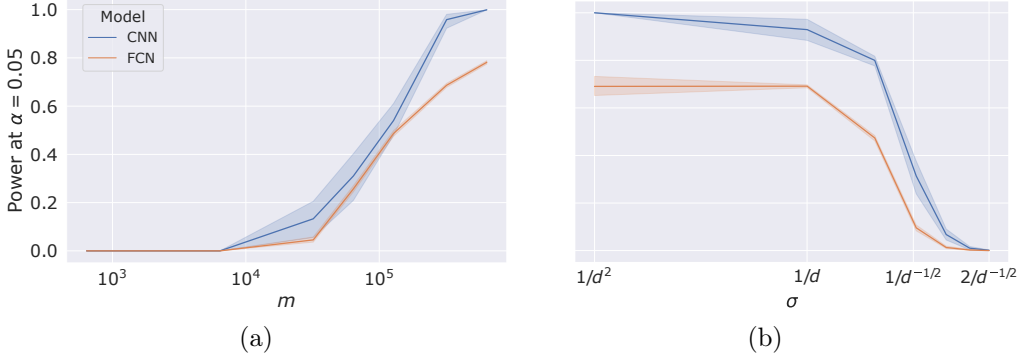
(a)   (b)

Figure 2: Estimation of the power of $1 - \gamma_{j,C}$ at level 0.05 for a simple CNN and a simple FCN on a synthetic dataset of images. (a) Displays power as a function of the number of samples m used for training, with fixed noise level $\sigma^2 = 1/d^2$, while (b) shows power as a function of the level of the noise in the test dataset while training on $m = 320 \times 10^3$ samples with fixed noise level $\sigma^2 = 1/d^2$.

for small values of α when all individual SHAPLIT tests for feature j are rejected, which suggests that this procedure has low power. In fact, $H_{0,j}^{\text{global}}$ is false as soon as at least one $H_{0,j,C}^{\text{SHAPLIT}}$ is false, but $\phi_j \geq 1 - \alpha/2$ only when *all* SHAPLIT nulls are rejected. Interesting future research directions entail fully characterizing the connection between the Shapley value and multiple hypothesis testing, and devising more powerful methods to combine p-values and test for the global null $H_{0,j}^{\text{global}}$.

4   Experiments

Here, we present three experiments that showcase how Theorem 2 can be used in practice.

4.1   Known Boolean function

We first study a case where both the distribution of the data and the ground-truth (Boolean) function are known. Let $k, n \in \mathbb{N}$ such that $\mathcal{X} \subseteq \mathbb{R}^{kn}$ and denote $X = [X_i, \ldots, X_k] \in \mathbb{R}^{kn}$ the concatenation of k vectors $X_i \in \mathbb{R}^n$. We here set $k = 2$, $n = 5$, and we define the ground-truth Boolean function

$$f(X) = (|(X_1)_1| \geq 3 \ \vee \ \ldots \ \vee \ |(X_1)_5| \geq 3) \bigwedge (|(X_2)_1| \geq 3 \ \vee \ \ldots \ \vee \ |(X_2)_5| \geq 3). \tag{14}$$

That is, the response $f(X)$ is positive if there is at least one feature above the threshold $t = 3$ in every vector $X_i \in X$. For an *important* distribution $\mathcal{I} = \mathcal{N}(4, 1)$ and its *unimportant* complement $\mathcal{I}^c = \mathcal{N}(0, 1)$, we sample the indices of the important features $(j^{(i)})_{i=1}^k$ uniformly over $[n]^k$, such that $(X_i)_{j^{(i)}} \sim \mathcal{I}$ and $(X_i)_{j \neq j^{(i)}} \sim \mathcal{I}^c$. Note that in this setting, every positive input has $k = 2$ important features with Shapley values of $1/k = 0.5$. In what sense should such features be considered important? Theorem 2 can be deployed to find which tests among the $2^{kn-1} = 512$ in $\phi_{j^{(i)}}(x, g)$ reject their null hypotheses. Fig. 1a shows that the bound in Theorem 2 is tight on empirical estimates of $p_{j,C}^{\text{SHAPLIT}}$ and $\gamma_{j,C}$ for the Shapley value of one important feature.[5] Fig. 1b displays the average empirical cumulative distribution of the estimates $\hat{p}_{j,C}^{\text{SHAPLIT}}$, $1 - \hat{\gamma}_{j,C}$ for all important features in 10 positive samples. Furthermore, it shows that half of the conditional independence tests in $\phi_{j^{(i)}}(x, g)$ reject their nulls at a critical value of 0.05, and that $1 - \hat{\gamma}_{j,C}$ correctly identifies them.

4.2   Synthetic image data

We now present a case where the distribution of the data is known, but we can only estimate the response through some learned model. Let $\mathcal{X} \subseteq \mathbb{R}^{dr \times ds}$, $d, r, s \in \mathbb{N}$ be images composed of an $r \times s$ grid of non-overlapping patches of $d \times d$ pixels, such that $X_{i,j} \in \mathbb{R}^{d \times d}$ is the patch in the ith-row and j^{th}-column. We consider a synthetic dataset of images where the response $y \in \{0, 1\}$ is positive if the input image $x \in \mathbb{R}^{dr \times ds}$ contains at least one instance of a target signal $x_0 \in \mathbb{R}^{d \times d}$. Denote images in $\mathbb{R}^{d \times d}$ as vectors in $\mathbb{R}^D$, $D = d^2$, and let $v \sim \mathcal{N}(0, \sigma^2 \cdot \mathbb{I}_D)$ be random noise. Define an *important* distribution $\mathcal{I} = x_0 + v$ for some target signal $x_0 \in \mathbb{R}^D$, and its *unimportant* complement $\mathcal{I}^c = v$, such that $X_{i,j} \sim a_{i,j} \cdot \mathcal{I} + (1 - a_{i,j}) \cdot \mathcal{I}^c$ and $a_{i,j} \sim \text{Bernoulli}(\eta)$ are independent Bernoulli random

---

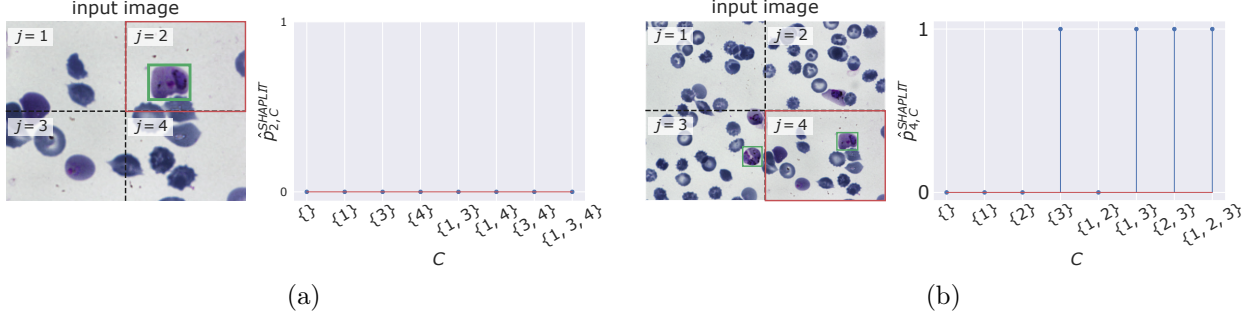[5]We set $K = 100 \times 10^3$ and $L = 1$ in the SHAPLIT procedure.

Figure 3: Demonstration of Shapley value (ϕ_j) and the p-values of their respective hypothesis tests, $p_{j,C}$, for two images of the BBBC041 dataset. Here, each feature is a quadrant, and important regions are those containing trophozoites (indicated by ground-truth green bounding box).

variables with parameter η. Then, $Y(X) = 1 \iff \exists (i,j) \in [r] \times [s] : X_{i,j} \sim \mathcal{I}$. In particular, we let $d = 7$, $r = s = 2$ and x_0 be a cross as in Fig. F.1a. Furthermore, we set $\eta = 1 - (1/2)^{1/(r \times s)}$ such that $\mathbb{P}[Y = 0] = \mathbb{P}[Y = 1] = 1/2$. Fig. F.1 shows some example images and their respective labels for different noise levels σ^2. We train a Convolutional Neural Network (CNN) and a Fully Connected Network (FCN) (see Appendix D.1 for further details). Recalling that the power of a test is defined as $1 - \beta = \mathbb{P}[\text{reject } H_0 \mid H_0 \text{ is false}]$, we estimate the power of performing conditional independence testing via Shapley coefficients at a chosen level α by evaluating $\mathbb{P}[1 - \gamma_{(i,j),C} \leq \alpha \mid H^{\text{SHAPLIT}}_{0,(i,j),C} \text{ is false}]$ on a test dataset of 320 samples. We remark that $H^{\text{SHAPLIT}}_{0,(i,j),C}$ is false for all patches $X_{i,j} \sim \mathcal{I}$ such that $X_{(i',j') \in C} \sim \mathcal{I}^c$, $C \subseteq [r] \times [s] \setminus \{i,j\}$. Fig. 2 shows an estimate of the power of $1 - \gamma_{(i,j),C}$ for both models as a function of number of samples shown during training (Fig. 2a), and noise in the test data (Fig. 2b) over 5 independent iterates.

4.3 Real image data

Finally, we revisit an experiment from [49] on the BBBC041 dataset [33], which comprises 1425 images of healthy and infected human blood smears of size 1200×1600 pixels. Here, the task is to label positively images that contain at least one trophozoite, an infectious type of cells. We apply transfer-learning to a ResNet18 [20] pretrained on ImageNet [12] (see Appendix D.2 for further details). After training, our model achieves a validation accuracy greater than 99%. For large dimensional inputs, as in this case, computing Shapley values for all players (i.e. all pixels) becomes intractable. Instead, we take an approach similar to that of *Hierarchical Shapley*, or h-Shap, in [49] and define sets of features as quadrants. These partitions can be applied hierarchically across different levels, provably guaranteeing a certain precision at every level in the hierarchy. This way, we extend the original implementation of h-Shap [49] to return the SHAPLIT tests that reject their respective nulls, or fail to reject them, thus assigning a collection of p-values to every feature (quadrant). Note that since there are 4 quadrants, each Shapley value (ϕ_j) comprises 8 terms ($\gamma_{j,C}$). Fig. 3 presents two examples of this setting. The first, Fig. 3a, depicts a case where there exist only 1 cell in the upper right quadrant, i.e. $j = 2$. Thus, this quadrant receives a Shapley value $\phi_2 \approx 1$. Naturally, all p-values are approximately zero (as also indicated by Corollary 1). This implies that the quadrant $j = 2$ is statistically important in the sense that all nulls are rejected. On the other hand, in Fig. 3b, there are two quadrants that contain cells, i.e. $j = 3, 4$. Hence, they are both important for this sample, and $\phi_3 \approx \phi_4 \approx 0.5$. Yet, based on our results, these can be decomposed into a sum of terms that bound the p-values of each of the tests performed by SHAPLIT. As can be seen, $j = 4$ is indeed statistically important in that half of the null hypotheses are rejected. We compare SHAPLIT with HRT [48] on retrieval of quadrants that contain at least one trophozoite in Appendix E. Overall, SHAPLIT increases both precision and f_1 score by 27% and 14% in absolute value over HRT, respectively.

5 Conclusion

The Shapley value and conditional hypothesis testing appear as two unrelated approaches to local interpretability (i.e. sample specific) of machine learning models. In this work, we have shown that

the two are tightly connected in that the former involves the computation of specific conditional hypothesis tests, and that every summand in the Shapley value can be used to bound the p-values of such tests. This perspective grants the Shapley value with a precise statistical meaning, allowing to control its Type I error. We present numerical experiments on synthetic and real data of increasing complexity to depict our theoretical results in practice. This work does not attempt to present a new method for interpretability. Instead, we intend to shed light on the meaning of game theoretic quantities in the context of statistical learning. These observations point to interesting open problems. For example—as for the Shapley value—some simplifying assumptions have to be made in order to overcome the computational complexity of carrying out an exponential number of tests, and the need to estimate the conditional distribution of high-dimensional data [34, 49, 8, 24]. Ideas related to restricting the Shapley value to an *a-priori* subset of null hypotheses may prove successful in devising useful algorithms in certain scenarios of practical relevance, where one may only be interested in a specific sense of feature importance. Furthermore, in Corollary 2, we suggest that the Shapley value itself bounds the p-value of a *global* null hypothesis [10] which is true if and only if all SHAPLIT nulls are true. Fully characterizing the connection between the Shapley value and multiple hypothesis testing is an exciting research direction. In particular, alternative ways other than the weighted arithmetic mean—which is employed by the Shapley value—may yield powerful procedures to combine SHAPLIT's p-values to test for global null hypotheses. These procedures will provide practitioners with simple and useful statements about a feature's importance. More generally, we hope the connection illustrated in this work will enable future developments on these issues, and, more importantly, will inform the responsible and safe deployment of explanation methods.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[2] Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.

[3] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271, 1997.

[4] Michael Bücker, Gero Szepannek, Alicja Gosiewska, and Przemyslaw Biecek. Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, 73(1):70–90, 2022.

[5] Collin Burns, Jesse Thomason, and Wesley Tansey. Interpreting black box models via hypothesis testing. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, pages 47–57, 2020.

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

[7] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*, 2018.

[8] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data. In *International Conference on Learning Representations*, 2019.

[9] Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

[10] Herbert A David and Haikady N Nagaraja. *Order statistics*. John Wiley & Sons, 2004.

[11] Lara Marie Demajo, Vince Vella, and Alexiei Dingli. Explainable ai for interpretable credit scoring. *arXiv preprint arXiv:2012.03749*, 2020.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017.

[14] Daniel Fryer, Inga Strümke, and Hien Nguyen. Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*, 9:144352–144360, 2021.

[15] Andreas Futschik, Thomas Taus, and Sonja Zehetmayer. An omnibus test for the global null hypothesis. *Statistical methods in medical research*, 28(8):2292–2304, 2019.

[16] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251. PMLR, 2019.

[17] Amirata Ghorbani, Michael Kim, and James Zou. A distributional framework for data valuation. In *International Conference on Machine Learning*, pages 3535–3544. PMLR, 2020.

[18] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[19] Philipp Hacker, Ralf Krestel, Stefan Grundmann, and Felix Naumann. Explainable ai under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law*, 28(4):415–439, 2020.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Nicholas A Heard and Patrick Rubin-Delanchy. Choosing between methods of combining-values. *Biometrika*, 105(1):239–246, 2018.

[22] Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007.

[23] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[24] Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-time shapley value estimation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Zq2G_VTV53T.

[25] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[26] Margot E Kaminski and Gianclaudio Malgieri. Algorithmic impact assessments under the gdpr: producing multi-layered explanations. *International Data Privacy Law*, pages 19–28, 2020.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. Cartoon explanations of image classifiers. *arXiv preprint arXiv:2110.03485*, 2021.

[29] Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. A rate-distortion framework for explaining black-box model decisions. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 91–115. Springer, 2022.

[30] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[31] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[32] Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and powerful conditional randomization testing via distillation. *arXiv preprint arXiv:2006.03980*, 2020.

[33] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012.

[34] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[35] Sisi Ma and Roshan Tourani. Predictive and causal implications of using shapley value for model interpretation. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, pages 23–38. PMLR, 2020.

[36] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[37] Art B Owen and Clémentine Prieur. On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002, 2017.

[38] Guillermo Owen. *Game theory*. Emerald Group Publishing, 2013.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

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

[42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[43] Juliet Popper Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46(1):561–584, 1995.

[44] Harshay Shah, Prateek Jain, and Praneeth Netrapalli. Do input gradients highlight discriminative features? *Advances in Neural Information Processing Systems*, 34, 2021.

[45] Lloyd S Shapley. *Notes on the N-person Game*. Rand Corporation, 1951.

[46] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551, 2021.

[47] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[48] Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M Blei. The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 31(1):151–162, 2022.

[49] Jacopo Teneggi, Alexandre Luster, and Jeremias Sulam. Fast hierarchical games for image explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–11, 2022. doi: 10.1109/TPAMI.2022.3189849.

[50] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*, 2018.

[51] Isabella Verdinelli and Larry Wasserman. Decorrelated variable importance. *arXiv preprint arXiv:2111.10853*, 2021.

[52] Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4): 791–808, 2020.

[53] Carlos Zednik. Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2):265–288, 2021.

[54] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

A Proofs

We briefly summarize the notation used in this section. Recall that $\hat{f} : \mathbb{R}^n \to [0,1]$ is a model trained on a set S of m i.i.d. samples from some distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{Y} = \{0,1\}$, and $x \sim \mathcal{D}_\mathcal{X} \in \mathbb{R}^n$ is a new sample. $\forall C \subseteq [n] := \{1, \ldots, n\}$ denote $M_C(x) = [x_C, \widetilde{X}_{-C}] \in \mathbb{R}^n$ the random vector that agrees with x in the features in C and that takes a reference value $\widetilde{X}_{-C} \sim X_{-C} \mid X_C = x_C$ in its complement.

A.1 Proof of Theorem 1

Here, we show that the p-value returned by the SHAPLIT procedure is valid, i.e. under $H_{0,j,C}^{\mathrm{SHAPLIT}}$, $\mathbb{P}[\hat{p}_{j,C}^{\mathrm{SHAPLIT}} \leq \alpha] \leq \alpha$, $\forall \alpha \in [0,1]$.

Proof. Recall that given a new sample $x \sim \mathcal{D}_\mathcal{X}$, $\forall C \subseteq [n]$, $\boldsymbol{M}_C(x) = (M_C^{(1)}(x), \ldots, M_C^{(L)}(x))$ are L masked random vectors, and $\widehat{\boldsymbol{Y}}_C = \hat{f}(\boldsymbol{M}_C(x))$ are the predictions of the model $\hat{f}$ on $\boldsymbol{M}_C(x)$. For a fixed feature $j \in [n]$ and $\forall C \subseteq [n] \setminus \{j\}$, the null hypothesis of the test is (see Eq. (6))

$$H_{0,j,C}^{\mathrm{SHAPLIT}} : \ \hat{f}\left(M_{C \cup \{j\}}(x)\right) \overset{d}{=} \hat{f}\left(M_C(x)\right). \tag{15}$$

It is easy to see that under the null hypothesis $H_{0,j,C}^{\mathrm{SHAPLIT}}$, the random vectors

$$\widehat{\boldsymbol{Y}}_{C \cup \{j\}}, \widehat{\boldsymbol{Y}}_C^{(1)}, \ldots, \widehat{\boldsymbol{Y}}_C^{(K)}, \tag{16}$$

$K \in \mathbb{N}$, are i.i.d. hence exchangeable. It follows that for any choice of test statistic $T(\widehat{\boldsymbol{Y}}_C)$, the random variables $T(\widehat{\boldsymbol{Y}}_{C \cup \{j\}}), T(\widehat{\boldsymbol{Y}}_C^{(1)}), \ldots, T(\widehat{\boldsymbol{Y}}_C^{(K)})$ are also exchangeable. Hence, $\mathbb{P}[\hat{p}_{j,C}^{\mathrm{SHAPLIT}} \leq \alpha] \leq \alpha$, $\forall \alpha \in [0,1]$. $\square$

A.2 Proof of Theorem 2

Here, we prove the bound on the p-value of the SHAPLIT procedure presented in Theorem 2 .

Proof. Given a feature $j \in [n]$, $\forall C \subseteq [n] \setminus \{j\}$ define $\Gamma_{j,C} := \hat{f}(M_{C \cup \{j\}}(x)) - \hat{f}(M_C(x))$. Assume that $\Gamma_{j,C}$ is a continuous random variable with probability density function $f_{\Gamma_{j,C}}(\gamma)$ (note that an equivalent version of this result can be provided for discrete random variables by considering their probability mass functions). Furthermore, let $T = \hat{f}(M_{C \cup \{j\}}(x))$, $\widetilde{T} = \hat{f}(M_C(x))$ be two distinct continuous random variables in $[0,1]$ with cumulative distribution functions $F_T(t)$, $F_{\widetilde{T}}(t)$, respectively, such that

$$F_T(t) = \mathbb{P}[T \leq t] \tag{17}$$
$$= \mathbb{P}[\hat{f}(M_{C \cup \{j\}}(x)) \leq t] \tag{18}$$
$$= \mathbb{P}\left[\widetilde{X}_{-(C \cup \{j\})} \in \widetilde{\mathcal{S}}_{t, C \cup \{j\}}\right], \tag{19}$$

and

$$F_{\widetilde{T}}(t) = \mathbb{P}[\widetilde{T} \leq t] \tag{20}$$
$$= \mathbb{P}[\hat{f}(M_C(x)) \leq t] \tag{21}$$
$$= \mathbb{P}\left[\widetilde{X}_{-C} \in \widetilde{\mathcal{S}}_{t, C}\right], \tag{22}$$

where $\widetilde{\mathcal{S}}_{t,C\cup\{j\}} := \{r \in \mathbb{R}^{n-|C\cup\{j\}|} : \hat{f}([x_{C\cup\{j\}}, r]) \leq t\} \subseteq \mathbb{R}^{n-|C\cup\{j\}|}$, and $\widetilde{\mathcal{S}}_{t,C} := \{r \in \mathbb{R}^{n-|C|} : \hat{f}([x_C, r]) \leq t\} \subseteq \mathbb{R}^{n-|C|}$ are two measurable sets. Let

$$p_{j,C} := \mathbb{P}\left[\Gamma_{j,C} \leq 0\right] = \int_{-\infty}^{0} \gamma f_{\Gamma_{j,C}}(\gamma) \, d\gamma, \tag{23}$$

$$\gamma_{j,C} := \mathbb{E}\left[\hat{f}(M_{C\cup\{j\}}(x))\right] - \mathbb{E}\left[\hat{f}(M_C(x))\right] \tag{24}$$

$$= \mathbb{E}[T] - \mathbb{E}\left[\widetilde{T}\right] \tag{25}$$

$$= \mathbb{E}\left[T - \widetilde{T}\right] \tag{26}$$

$$= \mathbb{E}\left[\Gamma_{j,C}\right], \tag{27}$$

where Eq. (26) follows from the known result that for any multivariate continuous random variable (X, Y), and any constants a, b, $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$, with $a = 1$, $b = -1$, $X = T$, and $Y = \widetilde{T}$. Note that $\Gamma_{j,C} \in [-1, 1]$ because difference of random variables in $[0, 1]$, then

$$\gamma_{j,C} = \mathbb{E}[\Gamma_{j,C}] = \int_{-\infty}^{\infty} \gamma f_{\Gamma_{j,C}}(\gamma) \, d\gamma = \int_{-1}^{1} \gamma f_{\Gamma_{j,C}}(\gamma) \, d\gamma \tag{28}$$

$$= \int_{-1}^{0} \gamma f_{\Gamma_{j,C}}(\gamma) \, d\gamma + \int_{0}^{1} \gamma f_{\Gamma_{j,C}}(\gamma) \, d\gamma \tag{29}$$

$$\leq \int_{0}^{1} \gamma f_{\Gamma_{j,C}}(\gamma) \, d\gamma \qquad\qquad (\int_{-1}^{0} \gamma f_{\Gamma_{j,C}}(\gamma) \, d\gamma \leq 0) \tag{30}$$

$$\leq \int_{0}^{1} f_{\Gamma_{j,C}}(\gamma) \, d\gamma \qquad\qquad (\gamma \leq 1) \tag{31}$$

$$= 1 - \int_{-1}^{0} f_{\Gamma_{j,C}}(\gamma) \, d\gamma \qquad\qquad (\int_{-1}^{1} f_{\Gamma_{j,C}}(\gamma) \, d\gamma = 1) \tag{32}$$

$$= 1 - p_{j,C}. \tag{33}$$

Rearranging the above inequality yields the statement of the theorem

$$p_{j,C} \leq 1 - \gamma_{j,C}. \tag{34}$$

$\square$

A.3 Proof of Corollary 1

Here, we provide the proof of the lower bound on $\gamma_{j,C}$ in $\phi_j(x, \hat{g})$, when $\phi_j(x, \hat{g}) \geq 1 - \epsilon$.

Proof. Given a feature $j \in [n]$, $\forall C \subseteq [n] \setminus \{j\}$ define $\Gamma_{j,C} := \hat{f}(M_{C\cup\{j\}}(x)) - \hat{f}(M_C(x))$ and denote $\hat{g}(M_C(x)) = \mathbb{E}[\hat{f}(M_C(x))]$ such that $\gamma_{j,C} = \mathbb{E}[\Gamma_{j,C}] = \mathbb{E}[\hat{f}(M_{C\cup\{j\}})(x))] - \mathbb{E}[\hat{f}(M_C(x))] \in [-1, 1]$. Recall that the tuple $(x, \hat{g})$, $x \in \mathbb{R}^n$ defines an n-person cooperative game such that its Shapley value can be rewritten as

$$\phi_j(x, \hat{g}) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \left(\hat{g}(M_{C\cup\{j\}}(x)) - \hat{g}(M_C(x))\right) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \gamma_{j,C}, \tag{35}$$

where $w_C = \binom{n-1}{|C|}^{-1} \cdot 1/n$ (see Eq. (9)). Assume that $\phi_j(x, \hat{g}) \geq 1 - \epsilon$, $\epsilon \in (0, 1)$ and fix $C^* \subseteq [n] \setminus \{j\}$ such that

$$\phi_j(x, \hat{g}) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \gamma_{j,C} = w_{C^*} \cdot \gamma_{j,C^*} + \sum_{C \neq C^*} w_C \cdot \gamma_{j,C} \geq 1 - \epsilon, \tag{36}$$

which implies

$$\gamma_{j,C^*} \geq \frac{1 - \epsilon - \sum_{C \neq C^*} w_C \cdot \gamma_{j,C}}{w_{C^*}} \tag{37}$$

$$\geq \frac{1 - \epsilon - \sum_{C \neq C^*} w_C}{w_{C^*}} \qquad (\max \gamma_{j,C} = 1) \tag{38}$$

$$= \frac{1 - \epsilon - (1 - w_{C^*})}{w_{C^*}} \qquad \left(\sum_{C \subseteq [n] \setminus \{j\}} w_C = 1\right) \tag{39}$$

$$= \frac{w_{C^*} - \epsilon}{w_{C^*}} \geq \frac{\widetilde{w} - \epsilon}{\widetilde{w}}, \quad \widetilde{w} = \min_{C \subseteq [n] \setminus \{j\}} w_C. \tag{40}$$

Note that the last inequality in Eq. (40) follows from $(w_{C^*} - \epsilon)/w_{C^*}$ being an increasing function of w_{C^*} for $w_{C^*} > 0$, which is always satisfied. Indeed, $\frac{d}{dw_{C^*}}\left((w_{C^*} - \epsilon)/w_{C^*}\right) = \epsilon/w_{C^*}^2 > 0$ when $w_{C^*} > 0$. Hence, to minimize γ_{j,C^*}, one needs to choose C^* such that $w_{C^*} = \widetilde{w} = \min_{C \subseteq [n] \setminus \{j\}} w_C$. Recall that $w_C = \binom{n-1}{|C|}^{-1} \cdot 1/n$. It follows that, for fixed n, $\min_{C \subseteq [n] \setminus \{j\}} \binom{n-1}{|C|}^{-1} = \binom{n-1}{\lfloor \frac{n-1}{2} \rfloor}^{-1}$ is the inverse of the central binomial coefficient and $\widetilde{w} = \binom{n-1}{\lfloor \frac{n-1}{2} \rfloor}^{-1} \cdot 1/n$. Finally, Stirling's approximation yields $\binom{n-1}{\lfloor \frac{n-1}{2} \rfloor}^{-1} = O(\sqrt{n}/4^{n/2})$, hence $\widetilde{w} = O(\sqrt{n}/n4^{n/2}) = O(1/\sqrt{n}4^{n/2})$. To conclude, note that there are 2^{n-1} many summands in $\phi_j(x, \hat{g})$, i.e. the number of tests in the Shapley value grows exponentially with n. To take into account the inflation of the Type I error due to multiple hypothesis testing, for example, apply the unweighted Bonferroni correction [43] and define $\alpha' = \alpha/2^{n-1}$, then

$$\min_{C \subseteq [n] \setminus \{j\}} \gamma_{j,C} \geq \frac{\widetilde{w} - \epsilon}{\widetilde{w}} \implies p_{j,C} \leq 1 - \frac{\widetilde{w} - \epsilon}{\widetilde{w}}, \quad \text{(from Theorem 2)} \tag{41}$$

and

$$p_{j,C} \leq 1 - \frac{\widetilde{w} - \epsilon}{\widetilde{w}} \leq \frac{\alpha}{2^{n-1}} \implies \epsilon \leq \widetilde{w} - \left(1 - \frac{\alpha}{2^{n-1}}\right) \cdot \widetilde{w} \tag{42}$$

$$= \frac{\alpha}{2^{n-1}} \cdot \widetilde{w} = O\left(\frac{\alpha}{2^{n-1}} \cdot \frac{1}{\sqrt{n}4^{n/2}}\right) \tag{43}$$

$$= O\left(\frac{\alpha}{4^{n/2}} \cdot \frac{1}{\sqrt{n}4^{n/2}}\right) = O\left(\frac{\alpha}{\sqrt{n}4^n}\right). \tag{44}$$

$\square$

A.4 Proof of Corollary 2

Here, we prove that the Shapley value of feature j can be used to test for a *global* null hypothesis $H_{0,j}^{\text{global}} = \bigcap_{C \subseteq [n] \setminus \{j\}} H_{0,j,C}^{\text{SHAPLIT}}$.

Proof. Given feature $j \in [n]$, denote $p_j^{\text{global}} = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot p_{j,C}$ the weighted arithmetic mean of the p-values of all SHAPLIT tests for feature j, where $w_C = \binom{n-1}{|C|}^{-1} \cdot 1/n$. Let ϕ_j be the Shapley

value of feature j. Then

$$\phi_j = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \gamma_{j,C} \tag{45}$$

$$\leq \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot (1 - p_{j,C}) \qquad \text{(from Theorem 2)} \tag{46}$$

$$= \sum_{C \subseteq [n] \setminus \{j\}} (w_C - w_C \cdot p_{j,C}) \tag{47}$$

$$= 1 - \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot p_{j,C} \qquad (\sum_{C \subseteq [n] \setminus \{j\}} w_C = 1) \tag{48}$$

$$= 1 - p_j^{\text{global}}. \tag{49}$$

Rearranging the above inequality yields the desired result

$$p_j^{\text{global}} \leq 1 - \phi_j. \tag{50}$$

We are left with showing that under $H_{0,j}^{\text{global}}$, $2 \cdot p_j^{\text{global}}$ is a valid p-value. To do so, we make use of Vovk and Wang [52, Proposition 9]. In particular, note that the weights w_C belong to the simplex $\Delta^{2^{n-1}-1}$, and that $1/w \coloneqq 1/\max_{C \subseteq [n] \setminus \{j\}} w_C = n \geq 2$ for $n \geq 2$, i.e. for more than two features. Then, setting $r = 1$ in the Proposition yields that two times the weighted arithmetic mean is a valid p-value. This concludes the proof of the corollary. $\qquad \square$

B Axioms of the Shapley value

Recall that the tuple $([n], v)$, $[n] \coloneqq \{1, \ldots, n\}$, $v : \mathcal{P}([n]) \to \mathbb{R}^+$ is an n-person TU game with characteristic function v, such that $\forall C \subseteq [n]$, $v(C)$ is the score accumulated by the players in the coalition C. Then, the Shapley values $\phi_1([n], v), \ldots, \phi_n([n], v)$ of the game $([n], v)$ (see Definition 1) are the only solution concept that satisfies the following axioms [45]:

Axiom 1 (Additivity)**.** The Shapley values sum up to the utility accumulated when all players participate in the game (i.e. the *grand coalition* of the game)

$$\sum_{j=1}^{n} \phi_j([n], v) = v([n]). \tag{51}$$

Axiom 2 (Nullity)**.** If a player does not contribute to any coalition, its Shapley value is 0

$$\forall C \subseteq [n] \setminus \{j\}, \ v(C \cup \{j\}) = v(C) \implies \phi_j([n], v) = 0. \tag{52}$$

Axiom 3 (Symmetry)**.** If the contributions of two players to any coalition are the same, their Shapley values are the same

$$\forall C \subseteq [n] \setminus \{j, k\}, \ v(C \cup \{j\}) = v(C \cup \{k\}) \implies \phi_j([n], v) = \phi_k([n], v). \tag{53}$$

Axiom 4 (Linearity)**.** Given $([n], v)$, $([m], v)$, the Shapley value of the union of the two games (i.e. $\phi_j([n] \cup [m], v)$) is equal to the sum of the Shapley values of the individual games (i.e. $\phi_j([n], v)$ and $\phi_j([m], v)$, respectively)

$$\phi_j([n] \cup [m], v) = \phi_j([n], v) + \phi_j([m], v). \tag{54}$$

Finally, we note that Axioms 2–4 can be replaced by a fifth one, usually referred to as *balanced contribution* [14], although this is not necessary to derive the definition of the Shapley value.

Algorithm C.1 Conditional Randomization Test

procedure CRT(data $\boldsymbol{X} = (x^{(1)}, \ldots, x^{(m)}) \in \mathbb{R}^{m \times n}$, response $\boldsymbol{Y} = (y^{(1)}, \ldots, y^{(m)}) \in \mathbb{R}^m$, feature $j \in [n]$, test statistic T, number of null draws $K \in \mathbb{N}$)

    Compute the test statistic, $t \leftarrow T(\boldsymbol{X}_j, \boldsymbol{Y}, \boldsymbol{X}_{-j})$

    for $k \leftarrow 1, \ldots, K$ **do**

        Sample $\widetilde{X}_j^{(i)} \sim X_j \mid X_{-j} = x_{-j}^{(i)}, \; i = 1, \ldots, m$

        $\widetilde{\boldsymbol{X}}_j \leftarrow (\widetilde{X}_j^{(1)}, \ldots, \widetilde{X}_j^{(m)})$

        Compute the null statistic, $\tilde{t}^{(k)} \leftarrow T(\widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \boldsymbol{Y})$

    end for

    return A (one-sided) p-value

$$\hat{p}_j^{\,\mathrm{CRT}} = \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbf{1}\left[\tilde{t}^{(k)} \geq t\right]\right)$$

end procedure

C   Algorithms

Algorithm C.1 summarizes the CRT procedure of Candes et al. [6].

D   Experimental details

Before describing the experimental details, we note that all experiments were run on an NVIDIA Quadro RTX 5000 with 16 GB of RAM memory on a private server with 96 CPU cores. All scripts were run on PyTorch `1.11.0` [39], Python `3.8.13`, and CUDA `10.2`.

D.1   Synthetic image data

Here, we describe the model architectures and the training details for the synthetic image datasets. Recall that $\mathcal{X} \subseteq \mathbb{R}^{dr \times ds}$, $d, r, s \in \mathbb{N}$ are images composed of an $r \times s$ grid of non-overlapping patches of $d \times d$ pixels, such that $X_{i,j} \in \mathbb{R}^{d \times d}$ is the patch in the ith-row and j^{th}-column.

We train a CNN with one filter with stride d, and a two-layer FCN with ReLU activation. In particular:

$$f^{\,\mathrm{CNN}}(X) = S\left(b_0 + \sum_{i,j \in [r] \times [s]} \langle W_0, X_{i,j} \rangle\right), \tag{55}$$

and

$$f^{\,\mathrm{FCN}}(X) = S\left(b_1 + \langle W_1, \mathrm{ReLU}\left(b_0 + \langle W_0, X \rangle\right)\rangle\right), \tag{56}$$

where $S(u) = 1/(1 + e^{-u})$ is the sigmoid function, and $\mathrm{ReLU}(u) = [0, x]_+$ is the rectified linear unit [36]. We train both models for one epoch on m i.i.d. samples and a batch size of 64. We note that we use Adam [27] with learning rate of 0.001, and SGD with learning rate of 0.01 for $f^{\,\mathrm{CNN}}$ and $f^{\,\mathrm{FCN}}$, respectively, to achieve optimal validation accuracy.

D.2   Real image data

Here, we present the details of the training process for the experiment on the BBBC041 dataset [33] (which is publicly available at https://bbbc.broadinstitute.org/BBBC041). Recall that the dataset comprises 1425 images of healthy and infected human blood smears. We split the original dataset into a training and validation split using an 80/20 ratio, respectively. This way, we train

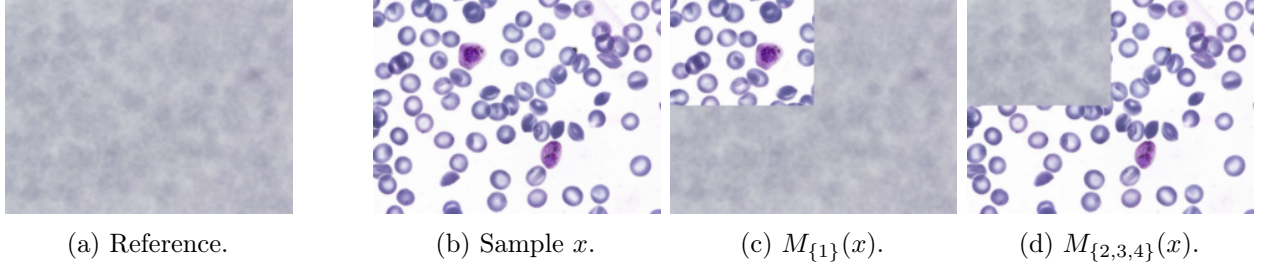|			
(a) Reference.	(b) Sample x.	(c) $M_{\{1\}}(x)$.	(d) $M_{\{2,3,4\}}(x)$.

Figure E.1: Some examples of masked inputs from the BBBC041 dataset. (a) Shows the unimportant reference (i.e. the unconditional expectation over the training split) used to mask features. We verify that the reference image does not contain any signal by checking that the output of $\hat{f}$ is ≈ 0. (b) Shows the original sample x; (c) and (d) show some masked versions of x when conditioning on the sets $C = \{1\}$ and $C = \{2, 3, 4\}$, respectively.

our model on 589 positive and 608 negative images, and validate on 112 positive and 116 negative images.

We apply transfer learning to a ResNet18 [20] pretrained on ImageNet [12]. We optimize all parameters of the network for 25 epochs using binary-cross entropy loss and Adam [27] optimizer, with a learning rate of 0.0001 and learning rate decay of 0.2 every 10 epochs. At training time, we augment the dataset with random horizontal flips.

E   Comparison with HRT

Here, we compare SHAPLIT and HRT [48] on the BBBC041 [33] dataset. Recall that HRT is a variation of the Conditional Randomization Test [6] that does not require to train several predictors. Instead, it uses the empirical risk of a fixed predictor $\hat{f}$ over holdout data as the test statistic. We remark that—in contrast with SHAPLIT—the null hypothesis of HRT is defined over a population rather than over a sample x. Given the different null hypotheses these two tests are designed for, we compare them in terms of detection of quadrants that contain at least one trophozoite rather than power. In particular, we label positively all quadrants that contain at least 80% of one trophozoite.[6] We compare tests in terms of precision and f_1 score over 101 predicted positive images. We note that, following [34, 49], we mask features with their unconditional expectation over the training split instead of estimating their conditional distribution. Fig. E.1 shows some example masked inputs on an image from the validation split of the BBBC041 dataset. Note that, since the masking process is deterministic, we set $K = 1$ in both tests and simply return the indicator $\hat{p} = \mathbf{1}[\tilde{t} \geq t]$ as p-value. We remark that this is suboptimal compared to having access to an estimate of the conditional distribution of the covariates. However, for simplicity, we assume that quadrants are almost independent. Finally, we threshold the original predictor $\hat{f}$ at 0.5—which minimizes the misclassification error—and use the binary predictor $\hat{f}_{01}(x) = \mathbf{1}[\hat{f}(x) \geq 0.5]$ as our model.

E.1   Detecting positive quadrants with HRT

Recall that the null hypothesis of HRT is $H_{0,j}^{\mathrm{HRT}} : \quad \hat{f} \perp\!\!\!\perp X_j \mid X_{-j}$. Here, we follow Algorithm 2 in Tansey et al. [48] and we use the 01 error over the validation split as the test statistic. We remark that, differently from SHAPLIT, HRT only tests for one null hypothesis for every feature $j \in [n]$, and it conditions on $C = -j = [n] \setminus \{j\}$ always, which corresponds to one of the 2^{n-1} subsets C in the SHAPLIT nulls for feature j. Hence, feature j is deemed important if $H_{0,j}^{\mathrm{HRT}}$ is rejected, i.e. if the 01 error increases when masking feature j. Denote t_{01} the 01 error on the validation split without masking any features, and $\tilde{t}_{01,j}$ the 01 error when masking feature j. We obtain $t = 0.44\%$, $\tilde{t}_{01,1} = 5.26\%$, $\tilde{t}_{01,2} = 2.63\%$, $\tilde{t}_{01,3} = 4.82\%$, and $\tilde{t}_{01,4} = 2.19\%$. Since masking any quadrant significantly increases the 01 error, all quadrants are deemed important by HRT, regardless of where trophozoites might be in a specific sample x. This result showcases one of the limitations

---

[6]We discard cases where trophozoites fall across quadrants given the partitioning strategy.

of population-level hypothesis tests when trying to explain a model's prediction on an individual sample.

E.2 Detecting positive quadrants with SHAPLIT

Recall from Corollary 2 that one can combine all 2^{n-1} SHAPLIT tests for feature j into one single valid p-value $p_j^{\text{global}} = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot p_{j,C}$, and use it to test for the *global* hypothesis $H_{0,j}^{\text{global}} : \bigcap_{C \subseteq [n] \setminus \{j\}} H_{0,j,C}^{\text{SHAPLIT}}$, which is true if and only if all $H_{0,j,C}^{\text{SHAPLIT}}$ are true. Intuitively, rejecting $H_{0,j}^{\text{global}}$ means that at least one SHAPLIT null is rejected for quadrant j, and we can conclude that j is important in some statistical sense for the prediction of the model, i.e. it should contain at least one trophozoite. Here—to account for the uncertainty of the model on real-world data—and similarly to [49], we use the following heuristic to deem a quadrant important. In particular, we say that feature j is important if its corresponding p_j^{global} is smaller than the 70th percentile of the global p-values of all 4 quadrants in the image. Differently from thresholding global p-values at a fixed level α, this approach allows for the threshold to be set relatively to the observed global p-values for a specific sample x.

E.3 Comparison

Table 1: Precision and f_1 scores of HRT and SHAPLIT on retrieval of positive quadrants on the BBBC041 dataset.

Test	precision	f_1
HRT	48%	64%
SHAPLIT	**75**%	**78**%

Table 1 reports precision and f_1 score for HRT and SHAPLIT for the retrieval of positive quadrants on the validation split of the BBBC041 dataset. Overall, SHAPLIT increases precision and f_1 score by 25% and 14%, respectively. This result showcases the advantage of deploying *local* conditional independence tests over a sample x rather than over a population when important features can vary in number (i.e. different images may contain different numbers of trophozoites) and position (i.e. trophozoites may appear in different regions of the image).

F Figures

(a) Signal x_0.

(b) $\sigma^2 = 1/d^4$.
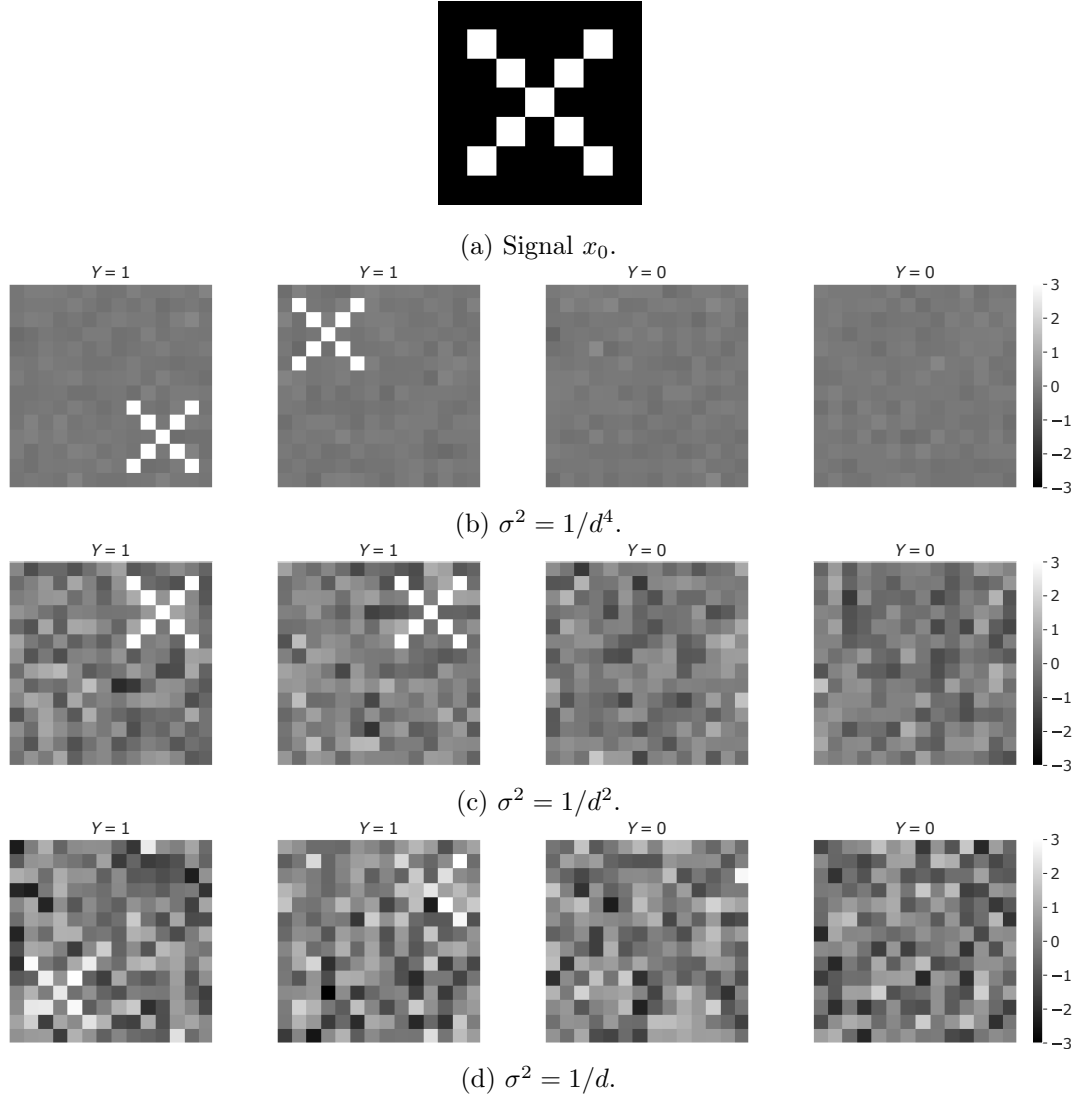

(c) $\sigma^2 = 1/d^2$.

(d) $\sigma^2 = 1/d$.

Figure F.1: (a) Target signal x_0: a cross of size $d \times d$ pixels, $d = 7$. (b), (c), (d) Examples of positive and negative images with increasing levels of noise, $\sigma^2 = 1/d^4$, $1/d^2$, and $1/d$, respectively. Note that all images are normalized to have zero-mean and unit variance.