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

Jacopo Teneggi\textsuperscript{1,3}, Beepul Bharti\textsuperscript{2,3}, Yaniv Romano\textsuperscript{4}, and Jeremias Sulam\textsuperscript{2,3}

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

October, 2022

Abstract

The complex nature of artificial neural networks raises concerns on their reliability, trustworthiness, and fairness in real-world scenarios. The Shapley value—a solution concept from game theory—is one of the most popular explanation methods for machine learning models. More traditionally, from the perspective of statistical learning, feature importance is defined in terms of conditional independence. So far, these two approaches to interpretability and feature importance have been considered separate and distinct. In this work, we show that Shapley-based explanation methods and conditional independence testing are closely related. We introduce the \textbf{SHAP}ley \textbf{L}ocal \textbf{I}ndependence \textbf{T}est (SHAPLIT), a novel testing procedure inspired by the Conditional Randomization Test (CRT) for a specific notion of local (i.e., on a sample) conditional independence. With it, we prove that for binary classification problems, each marginal contribution in the Shapley value is an upper bound to the $p$-value of this conditional independence test. Furthermore, we show that the Shapley value itself provides an upper bound to the $p$-value of a global SHAPLIT null hypothesis. As a result, we grant the Shapley value with a precise statistical sense of importance with false positive rate control.

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 [32]. These exciting results have come hand-in-hand with an increase in the complexity of these models, mostly based on neural networks. While these systems consistently set the state-of-the-art in many tasks, our understanding of their specific mechanisms remains intuitive at best. In fact, as neural networks keep getting deeper and wider, they also become \textit{opaque} (or unintelligible) to both developers and users. This lack of transparency is usually referred to as the \textit{black-box} problem: the predictions of a deep learning model are not readily interpretable, entailing theoretical, societal, and regulatory issues [65, 61, 56]. For example, it is still unclear whether a loss of interpretability is unavoidable to increase performance [49]. Furthermore, it may be required by certain regulations 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) [7, 28, 19].

These concerns on the reliability, trustworthiness, and fairness of machine learning have motivated recent efforts on explaining these opaque predictions. A distinction exists between \textit{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 [66], DeepLift [57], and LIME [48], the explainability literature has witnessed remarkable growth, and various approaches have been proposed to identify important features [50, 38, 8, 9, 30, 27, 31, 59, 41] with varying degrees of success [1].

In particular, the SHAP framework [38] unifies several existing methods while providing explanations that satisfy some desirable theoretical properties. More precisely, it brings the Shapley value [55]—a solution concept from game theory [45]—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 predictive model, where every feature in the input is a “player” in the game, and it computes their 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 [54]. The most important limitation of the Shapley value is its exponential computational cost in the number of players (i.e., the input dimension), which quickly becomes intractable for several applications of interest. Thus, most Shapley-based methods rely on various strategies to approximate them [38, 9, 26], some provably [59, 8].

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 the Shapley value to more traditional definitions of feature importance. Indeed, the statistics literature has a rich history of studying variable selection problems [3, 25, 18], for example, via conditional independence testing and controlled variable selection [6]. That is, a feature is unimportant if it is independent of a response 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, while assuming that the conditional distribution of the features is known instead. 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, 52, 51], or imaging data [43]). 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) by [58] and the Distilled Conditional Randomization Test (DCRT) by [34].

We remark that in this paper, we consider feature importance with respect to the response of a

---

1In a TU game, players can exchange their utility without incurring in any cost.
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 indeed important in many scenarios. For example, one may wish to understand the important features for a model that computes credit scores [10, 4], or one may need to verify that the important features for the prediction of an existing, complex model agree with prior-knowledge [5, 27, 13].

1.1 Related Work

Previous works have explored granting the Shapley value with some statistical notion of importance. For example, [44] and [62] study the Shapley value for correlated variables in the context of Analysis Of Variance (ANOVA) [24] and Leave Out COvariates (LOCO) [33] procedures, respectively. Importantly, [62] precisely raise the issue of the lack of statistical meaning for large (or small) Shapley values when features are correlated. [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 [39] try and deploy ideas of conditional independence to the Shapley value for causal inference on data generated by a Bayesian network. Our work differs from these, as we now summarize.

1.2 Contributions

In this paper, we will show that for 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 computed game-theoretic quantities provide an upper bound to the p-values of their respective tests. This novel connection provides several 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 and the power of the proposed tests.

We remark that we do not introduce a novel explanation method for model predictions. Our aim is to expand our fundamental understanding of the Shapley value when applied to machine learning models, particularly as its use for the interpretation of predictive models continues to grow [40, 67, 35]. These somewhat surprising insights should inform the design of novel explanation methods to guarantee 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., $x$) and random matrices with bold capital letters (e.g., $X$). Let $\mathcal{X}, \mathcal{Y}$ be some input and response domains, respectively, such that $(X,Y) \sim D$ is a labeled sample from a distribution $D$ 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 $D$ 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 $D^m$, with which one can choose a predictor $f$ that minimizes the empirical risk over $S$. In this work, we will not concern ourselves with this learning problem. Instead, we will assume we are given a fixed predictor, $f$, which we wish to analyze. More precisely, for a given input $x \sim D_{XY}$, we want to understand the importance of the features in $x$ with respect to the value of $f(x)$. This question will be formalized by means of the Shapley value, which we now define.
2.1 The Shapley Value

In game theory [45], the tuple \([n], v\) represents an \(n\)-person cooperative game with players \([n] := \{1, \ldots, n\}\) and characteristic function \(v : \mathcal{P}[n] \rightarrow \mathbb{R}^+\), 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 \(C \subseteq [n]\). Furthermore, one typically assumes \(v(\emptyset) = 0\). A Transfer Utility (TU) game is one where players can exchange their utility without incurring 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 [55] (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 of player \(j \in [n]\) is

\[
\phi_j([n], v) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot [v(C \cup \{j\}) - v(C)],
\]

where \(w_C = |C|! \cdot (n - |C| - 1)!/n!\). That is, \(\phi_j([n], v)\) is the average marginalized contribution of the \(j^{th}\) player over all subsets, or coalitions, of players.

2.2 Explaining Model Predictions with the Shapley Value

From Definition 1, the Shapley values 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 \(f(x)\) be the prediction of a learned model on a new sample \(x \sim \mathcal{D}_X\). 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). We will refer to

\[
\tilde{X}_C = [x_C, X_{-C}] \in \mathbb{R}^n
\]

as the random masked vector that is equal to \(x\) in the features in \(C\) and that takes an unimportant reference value in its complement. Following [38], we let \(X_{-C}\) be sampled from its conditional distribution given \(X_C = x_C\), i.e. \(X_{-C} \mid X_C = x_C\). Note that, in this way, \(X_{-C}\) is independent of the response \(f(x)\) by construction. For the sake of simplicity, we write \(\tilde{X}_C \sim \mathcal{D}_{X_C=x_C}\).

Given a predictor \(f\), denote \(v\) the characteristic function such that \(v(C) = g(\tilde{X}_C) = \mathbb{E} \left[ f(\tilde{X}_C) \right]\). This way, one can define the TU game \((x, g)\), where every feature in the sample \(x\) is a player in the game.\(^2\) Analogously to Definition 1, the Shapley value of feature \(j \in [n]\) in \((x, g)\) is

\[
\phi_j(x, g) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \left[ g(\tilde{X}_{C \cup \{j\}}) - g(\tilde{X}_C) \right].
\]

In the context of explainability, we look for a subset \(C^* \subseteq [n]\) such that \(x_{C^*}\) are the observed features that contributed the most towards \(f(x)\). For example, \(C^*\) may comprise the features with a large enough positive Shapley value, i.e. \(C^* = \{j \in [n] : \phi_j(x, g) > \tau\}\) for some \(\tau > 0\).

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. (3)) is that the model \(f\) has a fixed input domain \(\mathcal{X} \subseteq \mathbb{R}^n\), and furthermore it was trained on samples from a specific distribution over that domain, i.e. the marginal distribution \(\mathcal{D}_X\). Thus, one must mask features with a reference value \(X_{-C}\) from a specific conditional distribution so that \(\tilde{X}_C \sim \mathcal{D}_{X_C=x_C}\). Lastly, both the need to approximate the conditional distribution with high accuracy as well as the exponential number of summands in Eq. (3) make these Shapley values intractable in general. However, heuristic

\(^2(x, g)\) is with abuse of notation.
2.3 The Conditional Randomization Test (CRT)

We change gears to define variable importance from the perspective of conditional independence. Given some random variables $X, Y, Z$ we say that $X$ is independent of $Y$ given $Z$ (succinctly, $X \perp Y \mid Z$) if

$$
(X \mid Y, Z) \overset{d}{=} (X \mid Z),
$$

where $\overset{d}{=} \text{ indicates equality in distribution.}$

Recall that $\forall j \in [n], \; X_j \in \mathbb{R}$ is a random variable that corresponds to the $j^{th}$ feature in $X$, and $X_{-j} \in \mathbb{R}^{n-1}$ is its complement. Then, the CRT procedure proposed by [6] implements a conditional independence test for the null hypothesis

$$
H_{0,j}^{\text{CRT}} : \; X_j \perp \mid Y \mid X_{-j},
$$

which can be directly rewritten as

$$
H_{0,j}^{\text{CRT}} : \; (X_j \mid Y, X_{-j}) \overset{d}{=} (X_j \mid X_{-j}).
$$

Denote the random variable $X_{j}^{(k)} \sim X_j \mid X_{-j}$ a null duplicate of $X_j$ for $k = 1, \ldots, K$. We denote $X_{j}^{(k)}$ as null because, by construction, they are conditionally independent of $Y$ given $X_{-j}$. Hence, they come from the null distribution. Under $H_{0,j}^{\text{CRT}}$, for any choice of test statistic $T(X_j, Y, X_{-j})$, the random variables $T(X_j, Y, X_{-j}), T(X_j^{(1)}, Y, X_{-j}), \ldots, T(X_j^{(K)}, Y, X_{-j})$ are i.i.d., hence exchangeable [2]. Then, $\hat{p}_j^{\text{CRT}}$—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 $X = (X^{(1)}, \ldots, X^{(m)}) \in \mathbb{R}^{m \times n}, \; Y = (Y^{(1)}, \ldots, Y^{(m)}) \in \{0,1\}^m$ of $m$ i.i.d. random variables sampled from $D^m$.

3 HYPOTHESIS TESTING VIA THE SHAPLEY VALUE

Given the above background, we will now show that, somewhat surprisingly, the Shapley value is tightly connected with conditional independence testing. Formally, we present a novel modification of the CRT that, differently from the original motivation in Candès et al. [6], evaluates conditional independence of the response of a deterministic model $f$ on an individual sample $x \sim D_X$.

3.1 The SHAPley Local Independence Test (SHAPLIT)

Our SHAPLIT procedure tests for the independence of the response of a model $f$ on feature $x_j$ when the features in $x_C, \; C \subseteq [n] \setminus \{j\}$ are present, as we now formalize.

Definition 2 (SHAPley Local Independence Test). Let $\mathcal{X} \subseteq \mathbb{R}^n$, and $f : \mathcal{X} \rightarrow [0, 1]$ be a fixed given predictor for a binary response $Y \in \{0, 1\}$ on an input $X \in \mathcal{X}$ such that $(X,Y) \sim D$. Given a new sample $x \sim D_X$ and a subset of features $C \subseteq [n]$, denote $X_C = [x_C, X_{-C}]$ the corresponding random vector where $X_{-C}$ is sampled from its conditional distribution given $X_C = x_C$. We define the SHAPLIT null hypothesis as

$$
H_{0,j,C}^{\text{SHAPLIT}} : \; f(\tilde{X}_{C \cup \{j\}}) \overset{d}{=} f(\tilde{X}_C).
$$

Note that the SHAPLIT procedure can be generalized to any model $f : \mathcal{X} \rightarrow \mathcal{Y}$ for arbitrary domains $\mathcal{X}$ and $\mathcal{Y}$, but we will consider binary classification problems in this work. Similarly to the
Algorithm 1 Shapley Local Independence Test

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

Sample \( \tilde{X}_{C\cup\{j\}} \sim (\mathcal{D}_{X_{C\cup\{j\}}} = x_{C\cup\{j\}})^L \)
\( \tilde{Y}_{C\cup\{j\}} \leftarrow f(\tilde{X}_{C\cup\{j\}}) \)
Compute the test statistic \( t \leftarrow T(\tilde{Y}_{C\cup\{j\}}) \)
for \( k \leftarrow 1, \ldots, K \) do
Sample \( \tilde{X}_C^{(k)} \sim (\mathcal{D}_{X_C} = x_C)^L \)
\( \tilde{Y}_C^{(k)} \leftarrow f(\tilde{X}_C^{(k)}) \)
Compute the null statistic \( \tilde{p}^{(k)} \leftarrow T(\tilde{Y}_C^{(k)}) \)
end for
return A (one-sided) \( p \)-value
\[ \hat{p}_{j,C}^{\text{SHAPLIT}} = \frac{1}{K + 1} \left( 1 + \sum_{k=1}^{K} I[\tilde{p}^{(k)} \geq t] \right) \]
end procedure

HRT by [58] and the IRT by [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 by [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 \( j \)th one. Recalling from Section 2.2 that \( \tilde{X}_C \sim \mathcal{D}_{X_C} = x_C \) is a random masked vector equal to \( x \) in the features in \( C \), and that takes an unimportant (random) reference value—sampled from its conditional distribution—in its complement, one can rephrase the null hypothesis in Eq. (7) as

\[ \left( f(\tilde{X}_{C\cup\{j\}}) \mid X_{C\cup\{j\}} = x_{C\cup\{j\}} \right) \overset{d}{=} \left( f(\tilde{X}_C) \mid X_C = x_C \right). \]  

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 \( \tilde{X}_C = (\tilde{X}_C^{(1)}, \ldots, \tilde{X}_C^{(L)}) \in \mathbb{R}^{L \times n} \) the random matrix containing \( L \) i.i.d. samples from \( \mathcal{D}_{X_C} = x_C \), such that \( \tilde{Y}_C = f(\tilde{X}_C) \in \mathbb{R}^L \) are the predictions of \( f \) on each \( \tilde{X}_C^{(i)} \). For any choice of test statistic \( T(\tilde{Y}_C) \), 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 \( \hat{p}_{j,C}^{\text{SHAPLIT}} \)). Under the null hypothesis \( H_{\text{SHAPLIT}}^{\text{SHAPLIT}} \) in Eq. (7), \( \hat{p}_{j,C}^{\text{SHAPLIT}} \) the \( p \)-value returned by the SHAPLIT procedure—is valid for any choice of test statistic \( T \), i.e. \( \mathbb{P}[\hat{p}_{j,C}^{\text{SHAPLIT}} \leq \alpha] \leq \alpha \), \( \forall \alpha \in [0, 1] \).

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 \( H_{\text{SHAPLIT}}^{\text{SHAPLIT}} \), the distribution of the response of the model \( f \) is independent of \( x_j \) (the observed value of the \( j \)th feature in \( x \)) conditionally on \( X_C = x_C \). Since \( f \) is a deterministic function of its input, it is natural to ask when the random variable \( f(\tilde{X}_{C\cup\{j\}}) \) 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 \( f(\tilde{X}_{C\cup\{j\}}) \) is trivially degenerate. Then, when \( C = -j \) (i.e., all features but the \( j \)th one)
the complement set is empty (i.e., \(-(C \cup \{j\}) = \emptyset\)) and \(f(\tilde{X}_{C \cup \{j\}}(x))\) is point mass at \(f(x)\). For this setting, SHAPLIT retrieves the IRT procedure of [5] when testing only for the singleton set \(\{j\}\).

### 3.2 Connecting the Shapley Value and the SHAPLIT Conditional Independence Test

The attentive reader will have noticed that the Shapley value comprises the weighted sum of expected marginal contributions that are related to certain SHAPLIT null hypotheses. We now make this connection more precise.

To begin with, set \(L = 1\) in the SHAPLIT procedure and let \(T(\cdot)\) be the identity. Given a sample \(x \sim D_X\), a feature \(j \in [n]\), and a subset \(C \subseteq [n] \setminus \{j\}\), the test statistic is \(t = f(\tilde{X}_{C \cup \{j\}})\) and the null statistic is \(\tilde{t} = f(\tilde{X}_C)\). Furthermore, assume that \(f(x)\) is large for important features (otherwise, consider \(1 - f(x)\)), and denote

\[
\Gamma_{j,C} := f(\tilde{X}_{C \cup \{j\}}) - f(\tilde{X}_C)
\]

the random marginal contribution of feature \(j\) to the subset \(C\). Hence \(\gamma_{j,C} := \mathbb{E}[\Gamma_{j,C}] = \mathbb{E}[f(\tilde{X}_{C \cup \{j\}})] - \mathbb{E}[f(\tilde{X}_C)]\) and \(\lim_{K \to \infty} p_{j,C}^{\text{SHAPLIT}} = \mathbb{P}[\Gamma_{j,C} \leq 0]\). Furthermore, recall from Section 2.2 that \(g(\tilde{X}_C) = \mathbb{E}[f(\tilde{X}_C)]\), and thus the Shapley value of feature \(j\) can be rewritten as

\[
\phi_j(x,g) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \gamma_{j,C}.
\]

We now show that each summand \(\gamma_{j,C}\) is an upper bound to \(p_{j,C}\)—the \(p\)-value of its respective SHAPLIT test.

**Theorem 2.** Let \(X \subseteq \mathbb{R}^n\), and \(f : X \to [0,1]\) be a fixed given predictor for a binary response \(Y \in \{0,1\}\) on an input \(X \in X\) such that \((X,Y) \sim D\). Given a sample \(x \sim D_X\), a feature \(j \in [n]\), and a subset \(C \subseteq [n] \setminus \{j\}\), define \(\Gamma_{j,C} := f(\tilde{X}_{C \cup \{j\}}) - f(\tilde{X}_{C})\) such that \(\gamma_{j,C} = \mathbb{E}[\Gamma_{j,C}]\) and \(p_{j,C} = \mathbb{P}[\Gamma_{j,C} \leq 0]\). Then,

\[
p_{j,C} \leq 1 - \gamma_{j,C}.
\]

We defer the proof of this result to Appendix A.2, and provide now a few remarks.

First, note that while the bound above is on expectation, an analogous result can be easily derived by using finite-sample bounds by approximating \(\gamma_{j,C}\) with an empirical mean (as long as these are computed with samples independent of \(f\)).

Second, note that Theorem 2 provides a novel understanding of the Shapley value from a conditional independence testing perspective. In particular, each summand \(\gamma_{j,C}\) in \(\phi_j(x,g)\) is an upper bound to the \(p\)-value of the conditional independence test with null hypothesis \(H^{\text{SHAPLIT}}_{0,j,C}\).

Third, it is easy to see that \(\phi_j(x,g) = 1 \iff \gamma_{j,C} = 1\) for all \(C \subseteq [n] \setminus \{j\}\) concurrently. 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 \(\phi_j(x,g)\) reject their nulls. We now generalize this observation to more relevant cases where \(\phi_j(x,g) \geq 1 - \epsilon\), for small positive \(\epsilon\), by providing an upper bound to the \(p\)-values of all tests in \(\phi_j(x,g)\).

**Corollary 1.** In the setting of Theorem 2, if \(\phi_j(x,g) \geq 1 - \epsilon\), for \(0 \leq \epsilon \leq 1\), then \(\forall C \subseteq [n] \setminus \{j\}\),

\[
\gamma_{j,C} \geq \frac{\tilde{w} - \epsilon}{\tilde{w}},
\]

which implies

\[
p_{j,C} \leq 1 - \frac{\tilde{w} - \epsilon}{\tilde{w}},
\]

where \(\tilde{w} = \min_{C \subseteq [n] \setminus \{j\}} w_C\).
This result, whose proof we include in Appendix A.3, shows that a Shapley value \( \phi_j(x, g) \) close enough to 1 implies small \( p \)-values for all its SHAPLIT tests. Note, however, that \( \phi_j(x, g) \) comprises \( 2^{n-1} \) summands, hence \( \epsilon \) must decrease exponentially fast with \( n \) for Corollary 1 to be informative. Furthermore, to account for Type I error inflation due to multiple hypotheses testing, one can naturally include a correction as in [5]. For example, using the Bonferroni correction to control \( p_{j,C} \) at level \( \alpha \) yields \( \epsilon = O(\alpha/\sqrt{n4^n}) \).

Lasty, and conversely to the points above, note that these results also provide a new interpretation of a small 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 hypothesis, as the upper bound on the \( p \)-values will be uninformative. Similarly, we note that features with negative contributions, i.e. \( \gamma_{j,C} \leq 0 \), make the bound in Theorem 2 vacuous since \( p_{j,C} \leq 1 \) always. If one is interested in finding such features, it suffices to use \( t = 1 - f(\tilde{X}_{C \cup \{j\}}) \) and \( \tilde{t} = 1 - f(\tilde{X}_C) \) such that \( p_{j,C} = P[\Gamma_{j,C} \geq 0] \), and the analogous version of Theorem 2 follows. Recall that throughout this discussion, we have assumed \( f(x) \) to be large for important features (i.e., the presence of important features increases the model’s response). Intuitively, for binary classification problems, using \( 1 - f(x) \) as the test statistic is equivalent to explaining the opposite prediction, i.e. finding those features that lowered the model’s response.

### 3.3 The Shapley Value as a Global Hypothesis Test

The previous result linked each of the summands in the calculation of a Shapley coefficient with the \( p \)-value of a specific local hypothesis test. What could we say about the overall Shapley value? Such is the question we answer in this section.

An important task in multiple hypothesis testing is global testing. Given \( k \) null hypotheses \( H_{0,1}, \ldots, H_{0,k} \) with their respective \( p \)-values \( p_1, \ldots, p_k \), one may be interested in the overall null hypothesis \( H_0^{\text{global}} = \bigcap_{i=1}^k H_{0,i} \) which is true if and only if every individual null hypothesis is true [22]. This is by now a classical problem, and several procedures exist both based on the ordering of the \( p \)-values of the individual tests [23] and on ways of combining them [60, 47, 12]. More powerful alternatives have been recently proposed to overcome some of the limitations of these long-established methods, and, importantly, to make no assumptions on the dependency structure of the individual \( p \)-values [21, 15, 64, 63, 36]. We now show that the overall Shapley value of feature \( j \) (not just its individual summands \( \gamma_{j,C} \)) provides an upper bound to the \( p \)-value of a global SHAPLIT test.

Recall that for a feature \( j \in [n] \) there exists a SHAPLIT test with null hypothesis \( H_{0,j,C}^{\text{SHAPLIT}} \) for each \( C \subseteq [n] \setminus \{j\} \). Naturally, we can define the overall null hypothesis

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

which is false as soon as feature \( j \) is important with respect to at least one of these subsets \( C \subseteq [n] \setminus \{j\} \), and the following corollary of Theorem 2 follows.

**Corollary 2.** Denote \( p_j^{\text{global}} = 2 \cdot \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot p_{j,C} \), i.e. twice the weighted arithmetic mean of the \( p \)-values of all SHAPLIT tests for feature \( j \in [n] \). Then, under \( H_{0,j}^{\text{global}} \), \( p_j^{\text{global}} \) is a valid \( p \)-value and

\[
p_j^{\text{global}} \leq 2(1 - \phi_j(x, g)).
\]

The proof is included in Appendix A.4. A straightforward implication of this result is that, given a desired significance level \( \alpha \), one can reject \( H_{0,j}^{\text{global}} \) when \( \phi_j(x, g) \geq 1 - \alpha/2 \). However, recall from Corollary 1 that this condition implies that all individual SHAPLIT tests have exponentially

---

3See Appendix A.3 for details.

4Here, the term global carries a different meaning from that in the general machine learning explainability literature.
Figure 1: 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 (2 for each of the 10 samples considered). The black dotted vertical line marks the critical value of 0.05.

small $p$-values in $n$, which suggests that this procedure has low power. More precisely, $H_{0,j}^{\text{global}}$ is false as soon as one $H_{0,j,C}^{\text{SHAPLIT}}$ is false, but $\phi_j(x,g) \geq 1 - \alpha/2$ only when all SHAPLIT tests have small $p$-values. This highlights the suboptimality of the Shapley value from a general conditional independence testing perspective, which is specific to the way a global null is defined. At the same time, and importantly, Theorem 2 and Corollary 2 open the door to variations of the Shapley value inspired by more sophisticated ways of combining $p$-values other than averaging, which may prove more powerful in practice while guaranteeing false positive rate control. We consider these potential extensions as a part of future work.

4 EXPERIMENTS

We now present three experiments that showcase how the SHAPLIT procedure can be used in practice to explain machine learning predictions, contextualizing the Shapley value from a statistical viewpoint.

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 $X \subseteq \mathbb{R}^{kn}$ and denote $X = [X_1, \ldots, X_k] \in \mathbb{R}^{kn}$ the concatenation of $k$ vectors $X_i \in \mathbb{R}^n$. We define the ground-truth function $f(X) = \bigwedge_{i=1}^k D_i$ where $D_i = \bigvee_{j=1}^n (|(X_i)_j| \geq 3)$. In other words, the response 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 $I = \mathcal{N}(4,1)$ and its unimportant complement $I^c = \mathcal{N}(0,1)$, we sample the indices of the important features $(j^{(i)})_{i=1}^k$ uniformly over $[n]^k$, such that $(X_i)_{j^{(i)}} \sim I$ and $(X_i)_{j \neq j^{(i)}} \sim I^c$. In this setting, we set $k = 2$ important features in each input and $n = 5$. Hence, each important feature will receive a Shapley value of $1/k = 0.5$. Then: In what sense should such features be considered important from a statistical perspective? 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. In particular, Fig. 1 shows 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. We can see how 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. We present the distribution of estimates of $\hat{p}_{j,C}^{\text{SHAPLIT}}$ and $\hat{\gamma}_{j,C}$ for all 512 tests in the Shapley value of one important feature $j$ in Fig. F.1, which showcase the tightness of Theorem 2.
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 $X \subseteq \mathbb{R}^{dr \times ds}$, $d, r, s \in \mathbb{N}$ 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 $i^{th}$-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 variables with parameter $\eta$. 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 (an example is presented in Fig. F.2a). 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.2 shows some example images and their respective labels for different noise levels $\sigma^2$. We train a Convolutional Neural Network (CNN) and a Fully Connected Network (FCN) to predict the response $Y$ (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 $\alpha$ by evaluating $\mathbb{P}[1 - \gamma_{(i,j),C} \leq \alpha \mid H_0^{\text{SHAPLIT}} \text{ is false}]$ on a test dataset of 320 samples. We remark that $H_0^{\text{SHAPLIT}}$ 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.

### 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 $X \subseteq \mathbb{R}^{dr \times ds}$, $d, r, s \in \mathbb{N}$ 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 $i^{th}$-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 variables with parameter $\eta$. 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 (an example is presented in Fig. F.2a). 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.2 shows some example images and their respective labels for different noise levels $\sigma^2$. We train a Convolutional Neural Network (CNN) and a Fully Connected Network (FCN) to predict the response $Y$ (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 $\alpha$ by evaluating $\mathbb{P}[1 - \gamma_{(i,j),C} \leq \alpha \mid H_0^{\text{SHAPLIT}} \text{ is false}]$ on a test dataset of 320 samples. We remark that $H_0^{\text{SHAPLIT}}$ 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.

### Real Image Data

Finally, we revisit an experiment from [59] on the BBBC041 dataset [37], which comprises 1425 images of healthy and infected human blood smears of size $1200 \times 1600$ 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 [11] (see Appendix D.2 for further details). After training, our model achieves a validation accuracy greater than 99%. Since in this high-dimensional setting computing the exact Shapley value for each feature (i.e., each pixel) is intractable, we instead take an approach similar to that of Hierarchical Shapley, or $h$-Shap [59], 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-Shape 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). Since there are 4 quadrants, each Shapley value (φj) comprises only 8 terms (γj,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 φ2 ≈ 1. Naturally, all p-values are approximately zero (as also guaranteed 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 equally important for this sample and φ3 ≈ φ4 ≈ 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. However, the bound in Corollary 2 shows that a Shapley value of φ4 ≈ 0.5 is not large enough to reject the global null H0,global even if it is clearly false. Finally, and for the sake of completeness, we compare SHAPLIT with HRT [58] on retrieval of quadrants that contain at least one trophozoite in Appendix E. Note that for such quadrants, there exist at least one subset C ⊆ [n] \ {j} such that H0,global is false, hence the global hypothesis H0,global should be rejected. Overall, SHAPLIT increases both precision and f1 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. For the first time, this perspective grants the Shapley value with a precise statistical meaning, allowing to control its Type I error. We presented numerical experiments on synthetic and real data of increasing complexity to depict our theoretical results in practice. We hope that this work will enable the further and precise understanding of the meaning of (very popular) game theoretic quantities in the context of statistical learning. Restricting the Shapley value to an a-priori subset of null hypotheses may prove successful in devising useful algorithms in certain scenarios, and 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.
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] Bryan Casey, Ashkon Farhangi, and Roland Vogl. Rethinking explainable machines: The gdpr’s’ right to explanation’debate and the rise of algorithmic audits in enterprise. *Berkeley Tech. LJ*, 34:143, 2019.

[8] 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.

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

[11] 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.

[12] R. A. Fisher. Questions and answers. *The American Statistician*, 2(5):30–31, 1948. ISSN 00031305. URL http://www.jstor.org/stable/2681650.

[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] Gerhard Hommel. Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal*, 25(5):423–430, 1983.

[23] Gerhard Hommel, Frank Bretz, and Willi Maurer. Multiple hypotheses testing based on ordered p values—a historical survey with applications to medical research. *Journal of biopharmaceutical statistics*, 21(4):595–609, 2011.

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

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

[26] 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*, 2021.

[27] 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.

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

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

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

[31] 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.

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

[33] 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.
[34] Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 109(2):277–293, 2022.

[35] Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping, et al. Evaluating eligibility criteria of oncology trials using real-world data and ai. *Nature*, 592(7855):629–633, 2021.

[36] Yaowu Liu and Jun Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

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

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

[39] 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.

[40] Arturo Moncada-Torres, Marissa C van Maaren, Mathijs P Hendriks, Sabine Siesling, and Gijs Geleijnse. Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival. *Scientific Reports*, 11(1):1–13, 2021.

[41] Edoardo Mosca, Ferenc Szigeti, Stella Trangianni, Daniel Gallagher, and George Louis Groh. Shap-based explanation methods: A review for nlp interpretability. In *COLING*, 2022.

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

[43] Tuan-Binh Nguyen, Jérôme-Alexis Chevalier, and Bertrand Thirion. Ecko: Ensemble of clustered knockoffs for robust multivariate inference on fmri data. In *International Conference on Information Processing in Medical Imaging*, pages 454–466. Springer, 2019.

[44] 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.

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

[46] 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.

[47] Karl Pearson. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, pages 379–410, 1933.

[48] 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.

[49] 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.
[50] 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.

[51] Matteo Sesia and Tianshu Sun. Individualized conditional independence testing under model-x with heterogeneous samples and interactions. arXiv preprint arXiv:2205.08653, 2022.

[52] Matteo Sesia, Stephen Bates, Emmanuel Candès, Jonathan Marchini, and Chiara Sabatti. False discovery rate control in genome-wide association studies with population structure. Proceedings of the National Academy of Sciences, 118(40):e2105841118, 2021.

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

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

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] Leonard Henry Caleb Tippett et al. The methods of statistics. The Methods of Statistics., 1931.

[61] 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.

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

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

[64] Daniel J Wilson. The harmonic mean p-value for combining dependent tests. Proceedings of the National Academy of Sciences, 116(4):1195–1200, 2019.

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

[66] 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.
[67] Yazeed Zoabi, Shira Deri-Rozov, and Noam Shomron. Machine learning-based prediction of covid-19 diagnosis based on symptoms. npj digital medicine, 4(1):1–5, 2021.
A PROOFS

We briefly summarize the notation used in this section. Recall that \( f : \mathbb{R}^n \rightarrow [0, 1] \) 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 \( x \sim \mathcal{D}_X \in \mathbb{R}^n \) is a new sample. \( \forall C \subseteq [n] \) denote \( \hat{X}_C = [x_C, X_{-C}] \in \mathbb{R}^n \) the random masked vector that agrees with \( x \) in the features in \( C \) and that takes a reference value sampled from its conditional distribution \( X_{-C} | X_C = x_C \) in its complement. For brevity, we write \( \hat{X}_C \sim \mathcal{D}_{X_C=x_C} \).

A.1 Proof of Theorem 1

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

Proof. Recall that given a new sample \( x \sim \mathcal{D}_X \), \( \forall C \subseteq [n] \), \( \hat{X}_C \sim \mathcal{D}_{X_C=x_C} \) are \( L \) masked random vectors, and \( \hat{Y}_C = f(\hat{X}_C) \) are the predictions of the model \( f \) on \( \hat{X}_C \). For a fixed feature \( j \in [n] \) and \( \forall C \subseteq [n] \backslash \{j\} \), the null hypothesis of the test is (see Eq. (7))

\[
H_{0,j,C}^{\text{SHAPLIT}} : f(\hat{X}_{C \cup \{j\}}) \overset{d}{=} f(\hat{X}_C).
\]

(15)

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

\[
\hat{Y}_{C \cup \{j\}}, \hat{Y}_{C}^{(1)}, \ldots, \hat{Y}_{C}^{(K)},
\]

(16)

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

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 \( j \in [n] \) and a subset \( C \subseteq [n] \backslash \{j\} \) denote \( \Gamma_{j,C} := f(\hat{X}_{C \cup \{j\}}) - f(\hat{X}_C) \) the random marginal contribution of feature \( j \) with respects subset \( C \). 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 \( T = f(\hat{X}_{C \cup \{j\}}) \), \( \hat{T} = f(\hat{X}_C) \) be two distinct continuous random variables in \([0, 1]\) with cumulative distribution functions \( F_T(t) \) and \( F_{\hat{T}}(t) \), respectively. It follows that

\[
F_T(t) = \mathbb{P}[T \leq t]
\]

= \( \mathbb{P}[f(\hat{X}_{C \cup \{j\}}) \leq t] \)

= \( \mathbb{P}\left[X_{-(C \cup \{j\})} \in \tilde{S}_{t,C \cup \{j\}}\right] \),

(17)

(18)

(19)

and

\[
F_{\hat{T}}(t) = \mathbb{P}[\hat{T} \leq t]
\]

= \( \mathbb{P}[f(\hat{X}_C) \leq t] \)

= \( \mathbb{P}\left[X_{-C} \in \tilde{S}_{t,C}\right] \),

(20)

(21)

(22)
where \( \tilde{S}_{t,C \cup \{j\}} := \{ r \in \mathbb{R}^{n-|C \cup \{j\}|} : f([x_{C \cup \{j\}}, r]) \leq t \} \subseteq \mathbb{R}^{n-|C \cup \{j\}|} \), and \( \tilde{S}_{t,C} := \{ r \in \mathbb{R}^{n-|C|} : f([x_C, r]) \leq t \} \subseteq \mathbb{R}^{n-|C|} \) are two measurable sets. Let

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

\[
\gamma_{j,C} := \mathbb{E} \left[ f(\tilde{X}_{C \cup \{j\}}) - f(\tilde{X}_C) \right] = \mathbb{E}[T] - \mathbb{E}[\tilde{T}] \tag{24}
\]

\[
= \mathbb{E} [T - \tilde{T}] \tag{25}
\]

\[
= \mathbb{E} [\Gamma_{j,C}] , \tag{26}
\]

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

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

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

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

\[
\leq \int_{0}^{1} f_{\Gamma_{j,C}}(\gamma) \, d\gamma \quad (\gamma \leq 1) \tag{31}
\]

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

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

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 \( \phi_j(x, g) \), when \( \phi_j(x, g) \geq 1 - \epsilon \).

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

\[
\phi_j(x, g) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot (g(\tilde{X}_{C \cup \{j\}}) - g(\tilde{X}_C)) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \gamma_{j,C}, \tag{35}
\]

where \( w_C = \binom{n-1}{|C|}^{-1} \cdot 1/|n| \) (see Eq. (10)). Assume that \( \phi_j(x, g) \geq 1 - \epsilon, \epsilon \in (0, 1) \) and fix \( C^* \subseteq [n] \setminus \{j\} \) such that

\[
\phi_j(x, g) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \gamma_{j,C} = w_{C^*} \cdot \gamma_{j,C^*} + \sum_{C \neq C^*} w_C \cdot \gamma_{j,C} \geq 1 - \epsilon, \tag{36}
\]
which implies
\[
\gamma_{j,C^*} \geq \frac{1 - \epsilon - \sum_{C \neq C^*} w_C \cdot \gamma_{j,C}}{w_{C^*}} \geq \frac{1 - \epsilon - \sum_{C \neq C^*} w_C}{w_{C^*}} \geq \frac{1 - \epsilon - \left(1 - w_{C^*}\right)}{w_{C^*}} = \frac{\left(\sum\ C \subseteq [n] \setminus \{j\} w_C = 1\right)}{w_{C^*}} \quad \text{(max } \gamma_{j,C} = 1\text{)} \tag{38}
\]

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

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, \(\frac{\partial}{\partial w_{C^*}} ((w_{C^*} - \epsilon)/w_{C^*}) = \epsilon/w_{C^*}^2 > 0\) when \(w_{C^*} > 0\). Hence, to minimize \(\gamma_{j,C^*}\), one needs to choose \(C^*\) such that \(w_{C^*} = \tilde{w} = \min_{C \subseteq [n] \setminus \{j\}} w_C\).

Recall that \(w_C = \left(\frac{n-1}{|C|}\right)^{-1} \cdot 1/n\). It follows that, for fixed \(n\), \(\min_{C \subseteq [n] \setminus \{j\}} \left(\frac{n-1}{|C|}\right)^{-1} = \left(\frac{n-1}{2}\right)^{-1}\) is the inverse of the central binomial coefficient and \(\tilde{w} = \left(\frac{n-1}{\frac{n}{2}}\right)^{-1} \cdot 1/n\). Finally, Stirling’s approximation yields \(\left(\frac{n-1}{\frac{n}{2}}\right)^{-1} = O(\sqrt{n}/4^{n/2})\), hence \(\tilde{w} = O(\sqrt{n}/n4^{n/2}) = O(1/\sqrt{n}4^{n/2})\). To conclude, note that there are \(2^{2n-1}\) many summands in \(\phi_j(x, g)\), 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 [53] such that \(\alpha' = \alpha/2^{n-1}\). Then

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

and

\[
p_j,C \leq 1 - \frac{\tilde{w} - \epsilon}{\tilde{w}} \leq \frac{\alpha}{2n-1} \implies \epsilon \leq \tilde{w} - \left(1 - \frac{\alpha}{2n-1}\right) \cdot \tilde{w} \quad \text{(42)}
\]

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

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

\hfill \Box

### 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 \(p_j = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot p_{j,C}\) the weighted arithmetic mean of the \(p\)-values of all SHAPLIT tests for feature \(j\), where \(w_C = \left(\frac{n-1}{|C|}\right)^{-1} \cdot 1/n\). Let \(\phi_j\) be the Shapley value
of feature $j$. Then

$$\phi_j(x,g) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \gamma_{j,C}$$

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

(from Theorem 2)

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

$$= 1 - \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot p_{j,C}$$

(\sum_{C \subseteq [n] \setminus \{j\}} w_C = 1)

$$= 1 - p_j.$$ (49)

Rearranging the above inequality yields

$$p_j \leq 1 - \phi_j(x,g).$$ (50)

We are left with showing that under $H^{\text{global}}_{j,g}$, $p_j^{\text{global}} = 2p_j$ is a valid $p$-value. To do so, we make use of Vovk and Wang [63, Proposition 9]. In particular, note that the weights $w_C$ belong to the simplex $\Delta^{2^{n-1} - 1}$, and that $1/w = 1 / \max_{C \subseteq [n] \setminus \{j\}} w_C = n \geq 2$ 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. The statement of the corollary follows by multiplying each side of the inequality above by 2.

$$\square$$

## B AXIOMS OF THE SHAPLEY VALUE

Recall that the tuple $([n], v)$, $[n] := \{1, \ldots, n\}$, $v : \mathcal{P}([n]) \to \mathbb{R}^+$ 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 [55]:

**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)

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

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

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

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

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

$$\phi_j([n] \cup [m], v) = \phi_j([n], v) + \phi_j([m], v).$$ (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.
C ALGORITHMS

Algorithm C.1 summarizes the CRT procedure by [6].

Algorithm C.1 Conditional Randomization Test

\[
\text{procedure CRT(data } X = (x^{(1)}, \ldots, x^{(m)}) \in \mathbb{R}^{m \times n}, \text{response } Y = (y^{(1)}, \ldots, y^{(m)}) \in \mathbb{R}^m, \text{ feature } j \in [n], \text{ test statistic } T, \text{ number of null draws } K \in \mathbb{N}) \\
\text{Compute the test statistic, } t \leftarrow T(X_j, Y, X_{-j}) \\
\text{for } k \leftarrow 1, \ldots, K \text{ do} \\
\quad \text{Sample } X_j^{(k)} \sim X_j | X_{-j} = x_{-j}, i = 1, \ldots, m \\
\quad X_j^{(k)} \leftarrow (X_j^{(1)}, \ldots, X_j^{(m)}) \\
\quad \text{Compute the null statistic, } \tilde{t}^{(k)} \leftarrow T(X_j^{(k)}, X_{-j}, Y) \\
\text{end for} \\
\text{return A (one-sided) } p\text{-value } \hat{p}_{CRT}^j = \frac{1}{K+1} \left( 1 + \sum_{k=1}^K 1 \left[ \tilde{t}^{(k)} \geq t \right] \right) \\
\text{end procedure}
\]

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 [46], 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 \( X \subseteq \mathbb{R}^{dr \times ds} \), \( d, r, s \in \mathbb{N} \) 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 \( i^{th}\)-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:

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

and

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

where \( S(u) = 1/(1 + e^{-u}) \) is the sigmoid function, and \( \text{ReLU}(u) = [0, x]_+ \) is the rectified linear unit [42]. 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 [29] with learning rate of 0.001, and SGD with learning rate of 0.01 for \( f^{\text{CNN}} \) and \( f^{\text{FCN}} \), 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 [37] (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 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 [11]. We optimize all parameters of the network for 25 epochs using binary-cross entropy loss and Adam [29] 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 [58] on the BBBC041 [37] 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 $f$ 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. We compare tests in terms of precision and $f_1$ score over 101 predicted positive images. We note that, following [38, 59], 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 $\hat{p} = 1[\tilde{t} \geq t]$ 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 $f$ at 0.5—which minimizes the misclassification error—and use the binary predictor $f_{01}(x) = 1[f(x) \geq 0.5]$ as our model.

E.1 Detecting positive quadrants with HRT

Recall that the null hypothesis of HRT is $H_0^{\text{HRT}} : f \perp X_j | X_{-j}$. Here, we follow Algorithm 2 in Tansey et al. [58] 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 $H_0^{\text{HRT}}$ is

\footnote{We discard cases where trophozoites fall across quadrants given the partitioning strategy.}
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 \( t = 0.44\% \), \( t_{01,1} = 5.26\% \), \( t_{01,2} = 2.63\% \), \( t_{01,3} = 4.82\% \), and \( t_{01,4} = 2.19\% \). 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 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 \( p^\text{global}_j = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot p_{j,C} \), and use it to test for the global hypothesis \( H^\text{global}_j : \bigcap_{C \subseteq [n] \setminus \{j\}} H^\text{SHAPLIT}_{0,j,C} \), which is true if and only if all \( H^\text{SHAPLIT}_{0,j,C} \) are true. Intuitively, rejecting \( H^\text{global}_j \) 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 [59], we use the following heuristic to deem a quadrant important. In particular, we say that feature \( j \) is important if its corresponding \( p^\text{global}_j \) 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 \( \alpha \), 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).
Here we include supplementary figures.

Figure F.1: Estimates of $p_{\text{SHAPLIT}}^{j,C}$ and $\gamma_{j,C}$ for all 512 tests in the Shapley value of one important feature $j$ in the known Boolean setting.

Figure F.2: (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.