SHAP-XRT: The Shapley Value Meets Conditional Independence Testing

Anonymous authors
Paper under double-blind review

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 a statistical perspective, 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 **SHAP**ley-E**X**planation **R**andomization **T**est (SHAP-XRT), a 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, the marginal contributions in the Shapley value provide lower and upper bounds to the p-values of their respective tests. Furthermore, we show that the Shapley value itself provides an upper bound to the p-value of a *global* (i.e., overall) null hypothesis. As a result, we further our understanding of Shapley-based explanation methods from a novel perspective and characterize under which conditions one can make statistically valid claims about feature importance via the Shapley value.

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 (LeCun et al., 2015). 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 *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 readily interpretable, entailing theoretical, societal, and regulatory issues (Zednik, 2021; Tomsett et al., 2018; Shin, 2021). For example, it is still unclear whether a loss of interpretability is unavoidable to increase performance (Rudin, 2019). Furthermore, certain regulations may require companies and organizations who rely on autonomous systems to provide explanations of their decisions (e.g., what lead a model to reject a loan application) (Casey et al., 2019; Kaminski & Malgieri, 2020; Hacker et al., 2020).

These concerns on the reliability and trustworthiness of machine learning have motivated recent efforts on explaining 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 fashion (e.g., deep neural networks), we usually rely on *a-posteriori* explanation methods to find the most important features towards a prediction. These methods assign an importance score to every feature (or groups thereof) 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 (Zhou et al.,

2016), DeepLift (Shrikumar et al., 2017), and LIME (Ribeiro et al., 2016), the explainability literature has witnessed remarkable growth, and various approaches have been proposed to identify important features (Selvaraju et al., 2017; Lundberg & Lee, 2017; Chen et al., 2018; Covert et al., 2021; Kolek et al., 2021; Jiang et al., 2021; Kolek et al., 2022; Teneggi et al., 2022a; Mosca et al., 2022) with varying degrees of success (Adebayo et al., 2018).

In particular, the SHAP framework (Lundberg & Lee, 2017) unifies several existing methods while providing explanations that satisfy some desirable theoretical properties. More precisely, it brings the Shapley value (Shapley, 1951)—a *solution concept* from game theory (Owen, 2013)—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.[1] 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.[2] 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 (Shah et al., 2021). 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 (Lundberg & Lee, 2017; Covert et al., 2021; Jethani et al., 2021), some provably (Teneggi et al., 2022a; Chen et al., 2018).

While theoretically appealing, 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?

Previous works have explored the connection between feature selection and the Shapley value (Cohen et al., 2007; Fryer et al., 2021; Rozemberczki et al., 2022) as well as some statistical implications of large (or small) Shapley values (Owen & Prieur, 2017; Hooker, 2007; Ma & Tourani, 2020; Verdinelli & Wasserman, 2021; 2023). However, a general and precise connection between the Shapley value and statistical importance has been missing. In particular, it is unclear how to provide false positive rate control (i.e., to control the Type I error) when feature selection is performed on the basis of Shapley values. This is a fundamental issue as explanation methods are increasingly used in high-stakes scenarios to inform decision making, especially in medicine (Zoabi et al., 2021; de Almeida et al., 2022; Liu & Xie, 2020). For example, consider a clinical setting where a predictor is used to support diagnosis and treatment. Given each feature's Shapley value, which ones should we deem important? For example, if a patient's age, blood pressure, weight, and height receive Shapley values of 0.30, 0.15, 0.35, and 0.20, respectively, which should we report? It is easy to see how overlooking this question could lead to potentially harmful consequences in real-world scenarios.

In this paper, we show how relating the Shapley value to statistical notions of conditional independence provides a principled way to select important features. Indeed, the statistics literature has a rich history of studying variable selection problems (Blum & Langley, 1997; James et al., 2013; Guyon & Elisseeff, 2003), for example, via conditional independence testing and controlled variable selection (Candes et al., 2018). 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) (Candes et al., 2018) 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 (Candes et al., 2018; Sesia et al., 2021; Sesia & Sun, 2022), or imaging data (Nguyen et al., 2019)). While the original CRT procedure is computationally intractable for large models, fast and efficient alternatives have been recently proposed, such as the the

[1] In a TU game, players can exchange their utility without incurring in any cost.

[2] Herein, we will implicitly refer to the *Shapley value* of a feature according to the definition of SHAP attribution in Lundberg & Lee (2017).

Holdout Randomization Test (HRT) by Tansey et al. (2022) and the Distilled Conditional Randomization Test (DCRT) by Liu et al. (2022).

We remark that in this paper we consider 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 at the population level. 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 (Demajo et al., 2020; Bücker et al., 2022), or one may need to verify that the important features for the prediction of an existing, complex model agree with prior-knowledge (Burns et al., 2020; Jiang et al., 2021; Fong & Vedaldi, 2017).

1.1 Related work

Previous works have explored granting the Shapley value with some statistical notion of importance. For example, Owen & Prieur (2017) and Verdinelli & Wasserman (2021; 2023) study the Shapley value for correlated variables in the context of ANalysis Of VAriance (ANOVA) (Hooker, 2007) and Leave Out COvariates (LOCO) (Lei et al., 2018) procedures, respectively. Importantly, Verdinelli & Wasserman (2021) precisely raise the issue of the lack of statistical meaning for large (or small) Shapley values when features are correlated. Ghorbani et al. (2020) build upon Ghorbani & Zou (2019) and define the *Distributional Shapley* value to include information about the underlying data distribution into the original Shapley value. Recently, (Watson et al., 2023) propose a modified Shapley value with precise connections to information theory. Finally, and most closely to this work, Ma & Tourani (2020) 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 upper and lower bounds 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. We stress that we do not wish to introduce a novel explanation method for model predictions. Our aim is instead 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 (Moncada-Torres et al., 2021; Zoabi et al., 2021; Liu et al., 2021). These novel insights should inform the design of 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 and random vectors with capital letters (e.g., X), their realizations with lowercase letters (e.g., $X = x$) and random matrices with bold capital letters (e.g., $\boldsymbol{X}$). Let $\mathcal{X}, \mathcal{Y}$ be some input and response domains, respectively, such that $(X, Y) \sim \mathcal{D}$ is a random sample from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We set ourselves within the classical binary classification setting. Given $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{Y} = \{0, 1\}$, the task is to estimate the binary response Y on a sample X by means of a predictor $f : \mathcal{X} \to \mathcal{Y}$ that approximates the conditional expectation of the response on an input, i.e. $f(x) \approx \mathbb{E}[Y \mid X = x]$. We will not concern ourselves with this learning problem. Instead, we assume we are given a fixed predictor which we wish to analyze. More precisely, for an input $x \in \mathbb{R}^n$, 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 (Owen, 2013), the tuple $([n], v)$ represents an n-person cooperative game with players $[n] \coloneqq \{1, \ldots, n\}$ and characteristic function $v : \mathcal{P}([n]) \to \mathbb{R}^+$, where $\mathcal{P}(n)$ is the power set of $[n]$. For any subset of players $C \subseteq [n]$, the characteristic function outputs a nonnegative score $v(C)$ that represents the utility accumulated by the players in C. 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 (Shapley, 1951) (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

$$\phi_j([n], v) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \left[v(C \cup \{j\}) - v(C) \right], \tag{1}$$

where $w_C = |C|! \cdot (n - |C| - 1)!/n!$. That is, $\phi_j([n], v)$ is the average marginalized contribution of the jth player over all subsets (i.e., coalitions) of players.

2.2 Explaining model predictions with the Shapley value

Definition 1 shows how to compute the Shapley value 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 $x \in \mathbb{R}^n$. For any set of features $C \subseteq [n]$, denote $x_C \in \mathbb{R}^{|C|}$ the entries of x in the subset C (analogously, $x_{-C} \in \mathbb{R}^{n-|C|}$ for $-C \coloneqq [n] \setminus C$, the complement of C). We will refer to

$$\widetilde{X}_C = [x_C, X_{-C}] \in \mathbb{R}^n \tag{2}$$

as the random vector that is equal to x in the features in C and that takes a *random reference* value in its complement $-C$. Following Lundberg & Lee (2017), we let X_{-C} be sampled from its conditional distribution given x_C, i.e. $X_{-C} \mid X_C = x_C$.[3] Note that, in this way, X_{-C} is independent of the response $f(x)$ by construction. For the sake of simplicity, we write $\widetilde{X}_C \sim \mathcal{D}_{X_C = x_C}$. Then, for a predictor f and every subset $C \subset [n]$, $f(\widetilde{X}_C)$ is a random variable whose source of randomness is the random reference value.[4] This way, with abuse of notation, one can define the TU game (x, f), where every feature in the sample x is a player in the game. Therefore, analogous to Definition 1, the Shapley value of feature $j \in [n]$ is

$$\phi_j(x, f) = \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \mathbb{E}[f(\widetilde{X}_{C \cup \{j\}}) - f(\widetilde{X}_C)]. \tag{3}$$

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 $f(x)$.

We remark two main differences between the original definition of the Shapley value from game theory (Equation (1)) and its translation to machine learning models (Equation (3)) in common settings: (i) the characteristic function v can predict on sets of players, whereas the predictive model f typically has a fixed input domain $\mathcal{X} \subset \mathbb{R}^n$ (e.g., for a convolutional neural network);[5] and (ii) the characteristic function v is usually not data-dependent, while in this machine learning setting the model f was trained on samples from a specific distribution $\mathcal{D}$. As a result, given a subset of features $C \subseteq [n]$, one cannot simply predict on the partial input vector $x_C \in \mathbb{R}^{|C|}$ because $f(x_C)$ is not defined. Instead, the missing features in the complement of C must be masked with a reference value that does not contain any information about the response in

[3] Herein, we do not touch upon the discussion on *observational* vs. *interventional* Shapley values, and we refer the interested reader to (Sundararajan & Najmi, 2020; Chen et al., 2020; Merrick & Taly, 2020; Janzing et al., 2020).

[4] When $C = [n]$, $f(\widetilde{X}_C)$ is simply the prediction $f(x)$, which is not random.

[5] As noted in Jain et al. (2022); Covert et al. (2022); Teneggi et al. (2022b), Deep Sets (Zaheer et al., 2017) and Transformers (Vaswani et al., 2017) architectures do not suffer from this limitation.

order to simulate their absence. Furthermore, this reference value must come from the same data distribution $\mathcal{D}$ the model was trained on, hence the need to sample from the conditional distribution $X_{-C} \mid X_C = x_C$. Both the need to approximate the conditional distribution as well as the exponential number of summands in Equation (3) make the exact computation of Shapley values intractable in general. However, heuristic approximations have often provided useful results in practice (Lundberg & Lee, 2017), and in a few cases provable approximations also exist (Teneggi et al., 2022a; Chen et al., 2018).

2.3 The Conditional Randomization Test (CRT)

Before introducing our novel conditional independence testing procedure, we provide a brief summary of variable importance from the perspective of conditional independence. Given random variables X, Y, and Z we say that X is independent of Y given Z (succinctly, $X \perp\!\!\!\perp Y \mid Z$) if $(X \mid Y,\ Z) \overset{d}{=} (X \mid Z)$, where $\overset{d}{=}$ indicates equality in distribution. Put into words, knowing Y does not provide any more information about X in the presence of Z. Recall that for each $j \in [n],\ X_j \in \mathbb{R}$ is a random variable that corresponds to the jth feature in X, and $X_{-j} \in \mathbb{R}^{n-1},\ -j \coloneqq [n] \setminus \{j\}$ is its complement. Then, the CRT procedure by Candes et al. (2018) implements a conditional independence test for the null hypothesis

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

which can be directly rewritten as

$$H_{0,j}^{\text{CRT}} :\ (X_j \mid Y, X_{-j}) \overset{d}{=} (X_j \mid X_{-j}). \tag{5}$$

Let $X_j^{(k)} \sim X_j \mid X_{-j},\ k = 1, \dots, K$ be K *null* duplicates of X_j. $X_j^{(k)}$ are called nulls because, by construction, they are conditionally independent of the response Y given X_{-j}. Under $H_{0,j}^{\text{CRT}}$, for any choice of test statistic $T(X_j, Y, X_{-j})$, the random variables $T(X_j, Y, X_{-j}), T(X_j^{(1)}, Y, X_{-j}), \dots, T(X_j^{(K)}, Y, X_{-j})$ are i.i.d., hence exchangeable (Berrett et al., 2020). 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 Candes et al., 2018, Lemma F.1). For completeness, Algorithm C.1 summarizes the CRT procedure.

3 Hypothesis testing via the Shapley value

Given the above background, we will now show that the Shapley value for machine learning explainability (Equation (3)) is tightly connected with conditional independence testing. Formally, we present a novel modification of the CRT that, differently from the original motivation in Candes et al. (2018), evaluates conditional independence of the response of a deterministic model f on an individual sample $X = x,\ X \sim \mathcal{D}_{\mathcal{X}}$.

3.1 The Shapley Explanation Randomization Test (SHAP-XRT)

Going back to our motivating example of an automated system in a clinical setting, we would like to deploy ideas of conditional independence testing to report the most important features for the model's prediction on a patient. For example, we might be interested in knowing whether the response of the model on a specific patient is independent of age when their blood pressure and weight are collected. This way, we could provide precise statistical guarantees on the generated explanations, such as Type I error control. More precisely, this is equivalent to testing for the independence of the response of a model f on a feature $x_j,\ j \in [n]$ when the features in $x_C,\ C \subseteq [n] \setminus \{j\}$ are present (e.g., $j = \texttt{age}$ and $C = \{\texttt{blood pressure}, \texttt{weight}\}$). We now formalize this question by means of our novel SHAP-XRT null hypothesis.

Definition 2 (**SHAP**ley E**X**planation **R**andomization **T**est)**.** Let $\mathcal{X} \subset \mathbb{R}^n$ and $f :\ \mathcal{X} \to [0, 1]$ be a fixed predictor for a binary response $Y \in \{0, 1\}$ on an input $X \in \mathcal{X}$ such that $(X, Y) \sim \mathcal{D}$. Given a new sample $X = x,\ X \sim \mathcal{D}_{\mathcal{X}}$, a feature $j \in [n]$, and a subset of features $C \subseteq [n] \setminus \{j\}$, denote $\widetilde{X}_C = [x_C, X_{-C}]$ the corresponding random vector where X_{-C} is sampled from its conditional distribution given $X_C = x_C$. Then, the SHAP-XRT null hypothesis is

$$H_{0,j,C}^{\text{S-XRT}} :\ f(\widetilde{X}_{C \cup \{j\}}) \overset{d}{=} f(\widetilde{X}_C). \tag{6}$$

Algorithm 1 Shapley-Explanation Randomization Test (SHAP-XRT)

procedure SHAP-XRT(model $f: \mathbb{R}^n \to [0,1]$, 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 $\widetilde{\boldsymbol{X}}_{C\cup\{j\}} \sim \left(\mathcal{D}_{X_{C\cup\{j\}}=x_{C\cup\{j\}}}\right)^L$

$\widetilde{Y}_{C\cup\{j\}} \leftarrow f(\widetilde{\boldsymbol{X}}_{C\cup\{j\}})$

Compute the test statistic $t \leftarrow T(\widetilde{Y}_{C\cup\{j\}})$

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

Sample $\widetilde{\boldsymbol{X}}_C^{(k)} \sim \left(\mathcal{D}_{X_C=x_C}\right)^L$

$\widetilde{Y}_C^{(k)} \leftarrow f(\widetilde{\boldsymbol{X}}_C^{(k)})$

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

end for

return A (one-sided) p-value

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

end procedure

Recall that $\widetilde{X}_{C\cup\{j\}}$ and $\widetilde{X}_C$ are random vectors equal to x in the features in $C \cup \{j\}$ and C, respectively, and that take unimportant random reference values in their complements (Equation (2)). Then, the SHAP-XRT null hypothesis is equivalent to

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

which precisely asks whether the distribution of the response of the model changes when feature j is added to the subset C. We remark that the p-value returned by the SHAP-XRT procedure—akin to other conditional independence tests—is one-sided. Rejecting $H_{0,j,C}^{\text{S-XRT}}$ implies that in the presence of the features in C, the observed value of feature j is important in the sense that $f(\widetilde{X}_{C\cup\{j\}})$ is often *larger* (i.e., closer to 1), or stochastically greater, than $f(\widetilde{X}_C)$. If one desires to reject $H_{0,j,C}^{\text{S-XRT}}$ for features whose $f(\widetilde{X}_{C\cup\{j\}})$ is often *smaller* than $f(\widetilde{X}_C)$, it suffices to use $1 - f$ in lieu of f. Similarly, the SHAP-XRT procedure encompasses multi-class classification settings where $g: \mathbb{R}^d \to \Delta^M$ predicts one of M classes and $f(x) = g(x)_m,\ m \in [M]$. We stress that in contrast with the CRT null hypothesis (Equation (4)), $H_{0,j,C}^{\text{S-XRT}}$ is defined *locally* on a specific input x rather than over a population. Furthermore, $H_{0,j,C}^{\text{S-XRT}}$ can be generalized to any predictor $f: \mathcal{X} \to \mathcal{Y}$ for arbitrary domains, but we consider binary classification in this work. Finally, and differently from the Interpretability Randomization Test (IRT) by Burns et al. (2020), $H_{0,j,C}^{\text{S-XRT}}$: (i) asks for the conditional independence of the response with respect to a single feature $j \in [n]$ rather than a collection of subsets of features, and (ii) it allows to condition on arbitrary subsets $C \subseteq [n] \setminus \{j\}$ of features instead of restricting to $C = -j$, i.e. all features but j.

How should one test for this null hypothesis? Denote $\widetilde{\boldsymbol{X}}_{C\cup\{j\}} = (\widetilde{X}_{C\cup\{j\}}^{(1)}, \ldots, \widetilde{X}_{C\cup\{j\}}^{(L)}) \in \mathbb{R}^{L\times n}$ and $\widetilde{\boldsymbol{X}}_C = (\widetilde{X}_C^{(1)}, \ldots, \widetilde{X}_C^{(L)}) \in \mathbb{R}^{L\times n}$ the random matrices containing L duplicates of $\widetilde{X}_{C\cup\{j\}}$ and $\widetilde{X}_C$, respectively, such that $\widetilde{Y}_{C\cup\{j\}} = f(\widetilde{\boldsymbol{X}}_{C\cup\{j\}})$ and $\widetilde{Y}_C = f(\widetilde{\boldsymbol{X}}_C)$ are the predictions of the model. Algorithm 1 implements the SHAP-XRT testing procedure and it computes the respective p-value, $\hat{p}_{j,C}^{\text{S-XRT}}$, for any choice of test statistic T on the response of the model, e.g. the mean. We now formally state the validity of this test.

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

See Appendix A.1 for the proof of this result, and let us make a remark about the involved distributions. Recall that the SHAP-XRT procedure is defined locally on a sample x. Intuitively, under the null $H_{0,j,C}^{\text{S-XRT}}$, the distribution of the response of the model f is independent of x_j (the *observed* value of the jth 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 $f(\widetilde{X}_{C\cup\{j\}})$ has a degenerate distribution. Suppose that for each $C \subseteq [n] \setminus \{j\}$, $X_{-C} \mid X_C = x_C$ is not degenerate (e.g., it is not constant), otherwise $f(\widetilde{X}_{C\cup\{j\}})$ is trivially degenerate. For $C = -j$ (i.e., all features but the jth one), $\widetilde{X}_{C\cup\{j\}} = x$ because $C \cup \{j\} = [n]$ by definition of the complement set $-(C \cup \{j\})$ is empty, i.e. no reference value is sampled. It follows that for this choice of C, $f(\widetilde{X}_{C\cup\{j\}})$ is point mass at $f(x)$, and SHAP-XRT retrieves the IRT procedure (Burns et al., 2020).

3.2 Connecting the Shapley value and the SHAP-XRT conditional independence test

We now draw a precise connection between the SHAP-XRT testing procedure and the Shapley-value for machine learning explainability. This novel relation furthers our fundamental understanding of Shapley-based explanation methods, and, importantly, it clarifies under which conditions one can (and cannot) make statistically valid claims about feature importance through the Shapley value. We stress that—as we will discuss—this conditional independence interpretation is not intended to overcome the well-known computational limitations of the Shapley value in any way, nor to replace existing methods. Rather, it offers a previously overlooked perspective of such methods, and it suggests new strategies to develop more powerful, statistically well-grounded alternatives. To begin with, let T in the SHAP-XRT procedure be the identity and set $L = 1$ (we will discuss the effect of L on our results later). Then, for a sample x, a feature $j \in [n]$, and a subset $C \subseteq [n] \setminus \{j\}$, the SHAP-XRT test and null statistics become $t = f(\widetilde{X}_{C\cup\{j\}})$ and $\tilde{t} = f(\widetilde{X}_C)$, respectively. Denote

$$\Gamma_{j,C} \coloneqq f(\widetilde{X}_{C\cup\{j\}}) - f(\widetilde{X}_C) \tag{8}$$

the random marginal contribution of feature j to the subset C such that

$$p_{j,C}^{\text{S-XRT}} \coloneqq \mathbb{E}[\hat{p}_{j,C}^{\text{S-XRT}}] \qquad \text{and} \qquad \gamma_{j,C}^{\text{SHAP}} \coloneqq \mathbb{E}[\Gamma_{j,C}] \tag{9}$$

are the expected p-value of the SHAP-XRT procedure and the expected marginal contribution, respectively. Note the expectations are over both $\widetilde{X}_{C\cup\{j\}}$ and $\widetilde{X}_C$. This is to take into account the randomness in the draw of test statistic t. Then, it is easy to see that the Shapley value of feature j can be rewritten as

$$\phi_j(x, f) = \sum_{C\subseteq[n]\setminus\{j\}} w_C \cdot \gamma_{j,C}^{\text{SHAP}}. \tag{10}$$

We now show that each summand $\gamma_{j,C}^{\text{SHAP}}$ can be used to construct both upper and lower bounds to $p_{j,C}^{\text{S-XRT}}$.

Theorem 2. *Let $\mathcal{X} \subset \mathbb{R}^n$, and $f : \mathcal{X} \to [0, 1]$ be a fixed predictor for a binary response $Y \in \{0, 1\}$ on an input $X \in \mathcal{X}$ such that $(X, Y) \sim \mathcal{D}$. Given a sample $X = x$, $X \sim \mathcal{D}_{\mathcal{X}}$, a feature $j \in [n]$, and a subset $C \subseteq [n] \setminus \{j\}$, define $\Gamma_{j,C} \coloneqq f(\widetilde{X}_{C\cup\{j\}}) - f(\widetilde{X}_C)$ such that $p_{j,C}^{S\text{-}XRT} = \mathbb{E}[\hat{p}_{j,C}^{S\text{-}XRT}]$ and $\gamma_{j,C}^{SHAP} = \mathbb{E}[\Gamma_{j,C}]$. Then,*

$$0 \le p_{j,C}^{S\text{-}XRT} \le 1 - \frac{K}{K+1}\gamma_{j,C}^{SHAP} \tag{11}$$

and furthermore, if $\gamma_{j,C}^{SHAP} < 0$

$$\frac{1}{K+1}\left(1 + K(\gamma_{j,C}^{SHAP})^2\right) \le p_{j,C}^{S\text{-}XRT} \le 1, \tag{12}$$

where $K \ge 1$ is the number of draws of null statistic in the SHAP-XRT procedure.

We defer the proof of this result to Appendix A.2. We note that the theorem above is presented in expectation, and an analogous finite-sample result can be derived by means of concentration inequalities by estimating $\gamma_{j,C}^{\text{SHAP}}$ with an empirical mean (as long as it is computed over samples independent of f). We now discuss the meaning of Theorem 2 and its behavior as a function of K and L in the SHAP-XRT procedure, i.e. the number of draws of null statistic and the number of random reference values sampled at each iteration, respectively.

Behavior of Equations (11) and (12) as a function of K. As $K \to \infty$, i.e. as the number of draws of null statistic in the SHAP-XRT procedure goes to infinity, $K/(K+1) \to 1$, and Equations (11) and (12) become

$$0 \leq p_{j,C}^{\text{S-XRT}} \leq 1 - \gamma_{j,C}^{\text{SHAP}} \qquad \text{and} \qquad (\gamma_{j,C}^{\text{SHAP}})^2 \leq p_{j,C}^{\text{S-XRT}} \leq 1, \tag{13}$$

respectively. Furthermore, we remark that the coefficient $K/(K+1)$ in the bounds above is a mild requirement in practical scenarios. In particular, in order to expect to reject $H_{0,j,C}^{\text{S-XRT}}$ at level α for $\gamma_{j,C}^{\text{SHAP}} = 1$, it suffices for K to grow as $(1-\alpha)/\alpha$ (e.g., $\alpha = 0.05 \implies K \geq 19$).

Behavior of $\hat{p}^{\text{S-XRT}}$ as a function of L. Recall that L is the number of samples of reference values over which the test and null statistics are computed in the SHAP-XRT procedure. So far, we presented our results for $L = 1$ and T being the identity. These choices were instrumental for showcasing the connection between the SHAP-XRT test and the Shapley value. Here, we discuss the behavior of the SHAP-XRT p-value as L increases. First note that if $L > 1$, the test statistic T cannot simply be the identity because T is a function that maps $\mathbb{R}^L \to \mathbb{R}$. The SHAP-XRT procedure is valid for any choice of test statistic, T, and so for $L > 1$, choices of mean, median, quantile, max, etc, are all valid. For example, let T be the mean and in the limit $L \to \infty$, $t = \mathbb{E}[f(\widetilde{X}_{C\cup\{j\}})]$ and $\tilde{t} = \mathbb{E}[f(\widetilde{X}_C)]$. Then, $\hat{p}^{\text{S-XRT}} = (1/(K+1))^{\mathbf{1}[\gamma_{j,C}^{\text{SHAP}} > 0]}$.

Implications of Theorem 2 The result above provides a novel understanding of the Shapley value from a conditional independence testing perspective. In particular, it shows that the expected value of each summand $\gamma_{j,C}^{\text{SHAP}}$ (i.e., the expected marginal contribution of feature j to subset C) in the Shapley value of feature j provides both a lower and an upper bound to $p_{j,C}^{\text{S-XRT}}$, the expected p-value of its corresponding SHAP-XRT test (i.e., the test which asks whether the distribution of the response of the model is conditionally independent of j when the features in C are present). Until now, it was unclear whether the marginal contributions carried any statistical meaning at all. Theorem 2 precisely characterizes under which conditions one can (and cannot) use them to make statistically valid claims about feature importance. In particular, given a critical level α in the SHAP-XRT test, we can identify three separate conditions: (*i*) *large positive* contributions (i.e., $\gamma_{j,C}^{\text{SHAP}} > 1 - \alpha$) can provide sufficient evidence against the null hypothesis $H_{0,j,C}^{\text{S-XRT}}$, and so one can expect to reject it; (*ii*) *large negative* contributions (i.e., $\gamma_{j,C}^{\text{SHAP}} < -\sqrt{\alpha}$) cannot provide sufficient evidence against the null hypothesis $H_{0,j,C}^{\text{S-XRT}}$, and so one should not expect to be able to reject it; (*iii*) *small* contributions (i.e. $-\sqrt{\alpha} \leq \gamma_{j,C}^{\text{SHAP}} \leq 1 - \alpha$) cannot provide enough evidence against the null nor guarantee one should not expect to observe a p-value smaller than α. That is, one cannot expect to make any statistically valid claims. Figure 1 illustrates the bounds (when $K \to \infty$) and the three regimes described above.

3.3 The Shapley value as a global hypothesis test

The results presented so far study the statistical meaning of each summand (i.e., contribution) in the Shapley value. However, the overall Shapley value—and not just its individual summands—is used in practice to explain model predictions. Since every summand is linked to the p-value of a specific local hypothesis test, a naive approach would be to compare all p-values to level α, where rejecting at least one hypothesis implies feature j is important with respect to some subset $C \subseteq [n] \setminus \{j\}$. However, this approach is known to inflate the Type I error, and corrections are required (i.e., Bonferroni correction (Shaffer, 1995)). Instead, we ask whether the Shapley value is inherently linked to a *global* hypothesis test.[6] If so, what is the null hypothesis of this test? and what notion of importance does it convey? We now move onto answering these questions, for which we first introduce the definition of a global hypothesis test.

Given k null hypotheses $H_{0,1}, \ldots, H_{0,k}$ with their respective p-values $p_1, \ldots, p_k$, the global null hypothesis is defined as $H_0 = \bigcap_{i=1}^k H_{0,i}$ and it is true if and only if every individual null hypothesis is true (Hommel, 1983). Several classical procedures exist for global hypothesis testing, both based on the ordering of the p-values of the individual tests (Hommel et al., 2011) and on ways of combining them (Tippett et al., 1931; Pearson, 1933; Fisher, 1948). 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

[6]Here, the term *global* is used according to its meaning in the statistical theory literature, and not in the machine learning explainability literature.

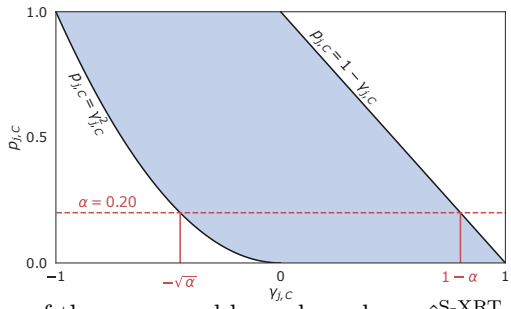


Figure 1: Pictorial representation of the upper and lower bounds on $\hat{p}_{j,C}^{\text{S-XRT}}$ as a function of $\gamma_{j,C}^{\text{SHAP}}$ described in Theorem 2 (as $K \to \infty$). For a critical level α (e.g., $\alpha = 0.20$ in the figure) the bounds indicate under what conditions one can expect to reject, or to fail to reject, the null hypothesis $H_{0,j,C}^{\text{S-XRT}}$ by means of $\gamma_{j,C}^{\text{SHAP}}$.

structure of the individual p-values (Heard & Rubin-Delanchy, 2018; Futschik et al., 2019; Wilson, 2019; Vovk & Wang, 2020; Liu & Xie, 2020).

With the above background, we now show how to aggregate all SHAP-XRT tests for feature $j \in [n]$ into a global SHAP-XRT null hypothesis test, provide a valid p-value for it, and finally highlight its relation to the Shapley value of feature j. Recall that there exists a SHAP-XRT test with null hypothesis $H_{0,j,C}^{\text{S-XRT}}$ for each $C \subseteq [n] \setminus \{j\}$. Using all these tests, we can define the global null hypothesis

$$H_{0,j} = \bigcap_{C \subseteq [n] \setminus \{j\}} H_{0,j,C}^{\text{S-XRT}}, \tag{14}$$

which is false as soon as one of the $H_{0,j,C}^{\text{S-XRT}}$ hypotheses is false. Put into words, rejecting $H_{0,j}$ implies that feature j is important (i.e., it has an effect on the distribution of the response of the model) with respect to at least one subset $C \subseteq [n] \setminus \{j\}$. This way, $H_{0,j}$ tests for a global notion of importance of feature j over all possible conditioning subsets. We remark that this is possible precisely because SHAP-XRT—differently from other conditional independence tests—allows to condition on arbitrary subsets of features. We now provide a valid p-value for this global test.

Lemma 1. *In the setting of Definition 2, and under the global null hypothesis $H_{0,j}$ (Equation (14)) the random variable*

$$\hat{p}_j = 2 \sum_{C \subseteq [n] \setminus \{j\}} w_C \cdot \hat{p}_{j,C}^{\text{S-XRT}} \tag{15}$$

is a valid p-value, i.e. $\mathbb{P}[\hat{p}_j \leq \alpha] \leq \alpha, \ \forall \alpha \in [0, 1]$.

This statement follows directly from (Vovk & Wang, 2020, Proposition 9) and its proof is included in Appendix A.3. We remark that the choice of weights is not unique as long as they belong to the simplex. Here, we use the same weights as the Shapley value's (i.e., uniform distribution over all possible subsets), as this will allow us to show that overall Shapley value of feature j indeed provides an upper bound to the expected p-value for the global SHAP-XRT test. Alternative weighting functions are subject of current research (Kwon & Zou, 2022) and we consider the study of their statistical meaning as future work.

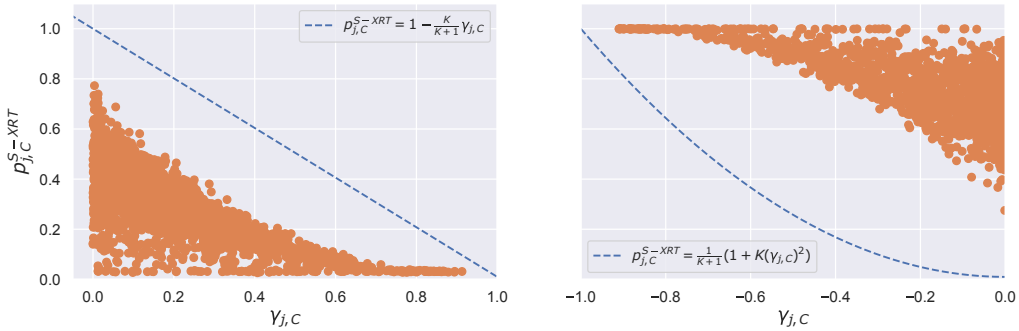
Corollary 1. *Denote $p_j = \mathbb{E}[\hat{p}_j]$ the expected p-value for the global null $H_{0,j}$. Then,*

$$p_j \leq 2 \left(1 - \frac{K}{K+1} \phi_j(x, f) \right), \tag{16}$$

where $\phi_j(x, f)$ is the Shapley value of feature $j \in [n]$.

The proof is included in Appendix A.4. This result shows that the Shapley value itself, and not only its summands, provides an answer to a statistical test. A straightforward implication of this result is that, given a desired significance level α, one can expect to reject $H_{0,j}$ when $\phi_j(x, f) \geq \frac{K+1}{K}(1 - \frac{\alpha}{2})$. In simple terms, a large positive Shapley value, $\phi_j(x, f) \approx 1$, implies that there exists *at least* one $C \subseteq [n] \setminus \{j\}$ for which its expected p-value, $p_{0,j,C}^{\text{S-XRT}}$ is small and $H_{0,j,C}^{\text{S-XRT}}$ is expected to be rejected. This result grants the Shapley value statistical meaning by demonstrating that it is linked to a global hypothesis test.

The natural questions that remain are then: (i) when $\phi_j(x, f) \approx 1$ which tests, amongst $\bigcap_{C \subseteq [n] \setminus \{j\}} H_{0,j,C}^{\text{S-XRT}}$, are being rejected? and (ii) similarly, do large negative Shapley values, $\phi_j(x, f) \approx -1$, carry any statistical meaning? We answer these questions with the following corollary in a similar fashion to Theorem 2.



(a) (b)

Figure 2: Demonstration of the upper and lower bounds of $p_{j,C}^{\text{S-XRT}}$ for the test $H_{0,j,C}^{\text{S-XRT}}: \; f(\widetilde{X}_{C\cup\{j\}}) \overset{d}{=} f(\widetilde{X}_C)$ with $j = 1, \; C = \{3, 4, 5, 6\}$ for the known response function experiment. (a) Displays the upper bound with $\theta_2 = 2$ for all samples with $\gamma_{j,C} \geq 0$. (b) Displays the lower bound with $\theta_2 = -2$ for all samples with $\gamma_{j,C} < 0$.

Corollary 2. *For a feature $j \in [n]$, if $\phi_j(x, f) \geq 1 - \epsilon, \; \epsilon \in (0, 1)$, then $\forall C \subseteq [n] \setminus \{j\}$*

$$p_{j,C}^{\text{S-XRT}} \leq 1 - \frac{K}{K+1}\frac{\widetilde{w} - \epsilon}{\widetilde{w}}, \tag{17}$$

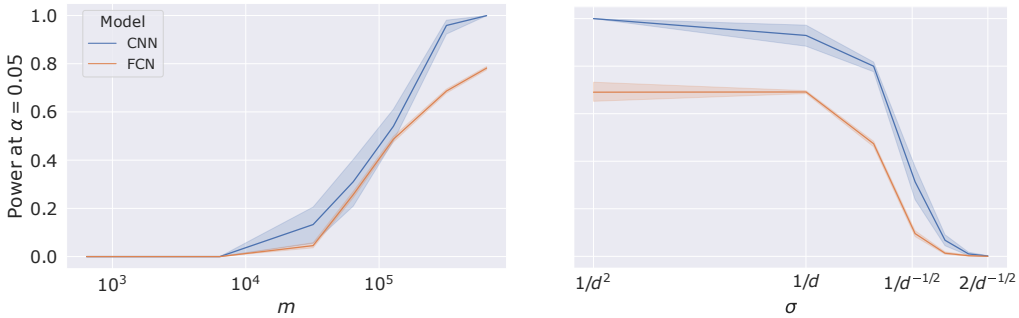
where $\widetilde{w} = \min_{C\subseteq[n]\setminus\{j\}} w_C$. Furthermore, if $\phi_j(x, f) \leq -1 + \epsilon$ with $\epsilon < \widetilde{w}$, then $\forall C \subseteq [n] \setminus \{j\}$

$$p_{j,C}^{\text{S-XRT}} \geq \frac{1}{K+1}\left(1 + K\left(\frac{\epsilon - \widetilde{w}}{\widetilde{w}}\right)^2\right). \tag{18}$$

The proof is included in Appendix A.5, and we now provide a few remarks. Firstly, Corollary 2 grants large positive (i.e, $\phi_j(x, f) \approx 1$) and large negative (i.e., $\phi_j(x, f) \approx -1$) Shapley values with statistical meaning. Recall from Corollary 1 that one can expect to reject the global null $H_{0,j}$ for large positive values of $\phi_j(x, f)$. This means that *at least one* of the SHAP-XRT null hypotheses of feature j is expected to be false. In other words, there exists at least one subset $C \subseteq [n] \setminus \{j\}$ such that j affects the distribution of the response of the model when the features in C are present. Corollary 2 makes this statement more precise. In fact, it shows that *all* SHAP-XRT tests needs to have small p-values, and one should expect to reject all tests. Symmetrically, it shows that for large negative Shapley values, one cannot expect to reject *any* SHAP-XRT null hypothesis for feature j. That is, feature j does not contribute to the distribution of the response of the model given any subset $C \subseteq [n] \setminus \{j\}$.

Secondly, Corollary 2 offers a statistical interpretation of the exponential cost of the Shapley value. Note that for the upper bound in Equation (17) not to be vacuous (i.e., less than 1) and for the lower bound in Equation (18) to hold (i.e., all contributions be negative), ϵ must be smaller than $\widetilde{w}$ (i.e., $\widetilde{w} - \epsilon > 0 \implies \epsilon < \widetilde{w}$ for Equation (17), and $\epsilon < \widetilde{w}$ for Equation (18) by construction). For n features, the weights w_C in the Shapley value assign a uniform distribution over all possible 2^{n-1} subsets $C \subseteq [n] \setminus \{j\}$. It follows that $\widetilde{w}$ is the inverse of the central binomial coefficient, i.e. $\widetilde{w} = \frac{1}{n} \cdot \binom{n-1}{\lfloor \frac{n-1}{2} \rfloor}^{-1} = \mathcal{O}(\frac{\sqrt{n}}{n4^{n/2}})$ by Stirling's approximation. Then, ϵ needs to decay *exponentially fast with* n for the bounds to be informative. This is a well-known limitation of the Shapley value in practical high-dimensional settings. We remark that studying under which settings the computational cost of the Shapley value can be reduced is subject of ongoing research that is orthogonal to the contribution of this paper (Chen et al., 2019; Teneggi et al., 2022a). Rather, Corollary 2 shows the statistical implications of the exponential number of subsets one needs to account for when computing the Shapley value of feature j.

Finally, Corollaries 1 and 2 open the door to variations of the Shapley value inspired by more sophisticated and powerful ways of combining p-values other than averaging, which may provide more effective in practice while guaranteeing false positive rate control. We consider these potential extensions as part of future work.



(a) (b)

Figure 3: 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 $\sigma^2 = 1/d^2$. We set $K = 100 \times 10^3$ and $L = 1$ in the SHAPLIT procedure.

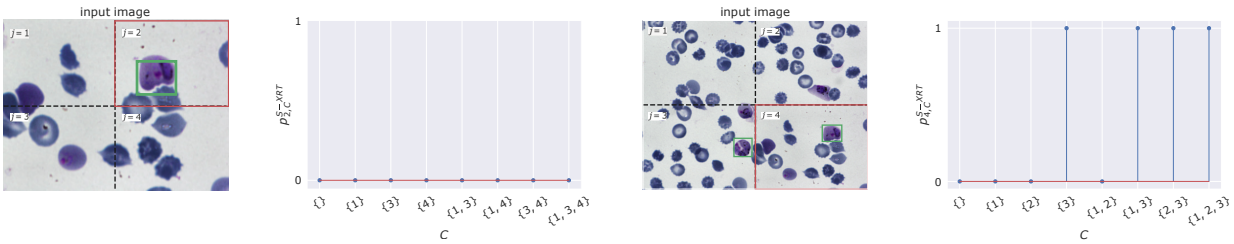
4 Experiments

We now present three experiments, of increasing complexity, that showcase how the SHAP-XRT procedure can be used in practice to explain machine learning predictions, contextualizing the Shapley value from a statistical viewpoint. All code to reproduce experiments will be made publicly available.

Known response function We first study a case where both the distribution of the data and the ground-truth function are known. Let $d \in \mathbb{N}$ such that $\mathcal{X} \subseteq \mathbb{R}^{2d}$ and denote $X = [X_1, \ldots, X_d] \in \mathbb{R}^{2d}$ the concatenation of d vectors $X_i = [X_{i,1}, X_{i,2}] \in \mathbb{R}^2$. We define the ground-truth function $f(X) = S(\theta^T X)$ where S is the sigmoid function[7] and $\theta = [\theta_1, \theta_2, \cdots, \theta_{2d}] \in \mathbb{R}^{2d}$. For $i \in [d]$, $X_{i,1} \sim \mathcal{N}(1,1)$ and $X_{i,2} \sim \mathcal{N}(3, \frac{1}{4})$ if $X_{i,1} > 3$ and $\mathcal{N}(-1,1)$ otherwise. We set $d = 3$ and let $\theta = \mathbf{1}_{2d}$ with the exception that $\theta_2 = 2$. We concern ourselves with the null hypothesis $H_{0,j,C}^{\text{S-XRT}} : \; f(\widetilde{X}_{C\cup\{j\}}) \overset{d}{=} f(\widetilde{X}_C)$ with $j = 2$ and $C = \{3,4,5,6\}$. That is, we are aiming to detect when feature $j = 2$, in the presence of the features in $C = \{3,4,5,6\}$, increases the value of $f(\widetilde{X}_{C\cup\{j\}})$ relative to $f(\widetilde{X}_C)$. Using the SHAP-XRT procedure with $K = 100$, we calculate $p_{j,C}^{\text{S-XRT}}$ and $\gamma_{j,C}$ for 1000 samples. In Figure 2a, for all samples with $\gamma_{j,C} \geq 0$, we plot $p_{j,C}^{\text{S-XRT}}$ as a function of $\gamma_{j,C}$ and display the upper bound $p_{j,C}^{\text{S-XRT}} = 1 - \frac{K}{K+1}\gamma_{j,C}$. To demonstrate the validity of the lower bound, we keep the experiment the same except that now $\theta_2 = -2$. In Figure 2b, for all samples with $\gamma_{j,C} < 0$, we plot $p_{j,C}^{\text{S-XRT}}$ as a function of $\gamma_{j,C}$ and display the lower bound $p_{j,C}^{\text{S-XRT}} = \frac{1}{K+1}\left(1 + K(\gamma_{j,C})^2\right)$.

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 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 (an example is presented in Figure E.2a). Furthermore, we set $\eta = 1 - (1/2)^{1/(r \times s)}$ such that $\mathbb{P}[Y = 0] = \mathbb{P}[Y = 1] = 1/2$. Figure E.2 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) 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 α by evaluating $\mathbb{P}[1 - \gamma_{(i,j),C} \leq \alpha \mid H_{0,(i,j),C}^{\text{S-XRT}} \text{ is false}]$ on a test dataset of 320 samples. We remark that $H_{0,(i,j),C}^{\text{S-XRT}}$ is false for all patches $X_{i,j} \sim \mathcal{I}$ such that $X_{(i',j') \in C} \sim \mathcal{I}^c$, $C \subseteq [r] \times [s] \setminus \{i,j\}$. Figure 3 shows an estimate of the power of $1 - \gamma_{(i,j),C}$ for both models

[7] $S(u) = 1/(1 + e^{-u})$.



(a) (b)

Figure 4: 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).

as a function of number of samples shown during training (Figure 3a), and noise in the test data (Figure 3b) over 5 independent training realizations.

Real image data Finally, we revisit an experiment from Teneggi et al. (2022a) on the BBBC041 dataset (Ljosa et al., 2012), 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 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) (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 (Teneggi et al., 2022a), 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 to return the SHAP-XRT tests that reject their respective nulls, or fail to reject them, thus assigning a collection of p-values to every feature (quadrant). We note that in this experiment, we use the *unconditional* expectation over the training set to mask features—as it is commonly done in popular Shapley-based explanation methods. Figure E.1 shows the reference value used for masking and some example masked inputs. Since there are 4 quadrants, each Shapley value (ϕ_j) comprises only 8 terms ($\gamma_{j,C}$). Figure 4 presents two examples of this setting. The first, Figure 4a, 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 guaranteed by Corollary 2). This implies that the quadrant $j = 2$ is statistically important in the sense that all nulls are rejected. On the other hand, in Figure 4b, there are two quadrants that contain cells, i.e. $j = 3, 4$. Hence they are both equally 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 SHAP-XRT. As can be seen, $j = 4$ is indeed statistically important in that half of the null hypotheses are rejected. However, the bound in Corollary 1 shows that a Shapley value of $\phi_4 \approx 0.5$ is not large enough to reject the global null $H_{0,j}$ even if it is clearly false. The results in this experiment showcase the suboptimality of the Shapley value from a statistical testing perspective, and motivate future work to develop more powerful alternatives.

5 Conclusion

The Shapley value and conditional hypothesis testing appear as two unrelated approaches to local (i.e., sample specific) interpretability 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. 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 SHAP-XRT's p-values to test for global null hypotheses.

References

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.

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.

Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The american mathematical monthly*, 107(4):353–357, 2000.

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

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.

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*, pp. 47–57, 2020.

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.

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.

Hugh Chen, Joseph D Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

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.

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.

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

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.

Ian Covert, Chanwoo Kim, and Su-In Lee. Learning to estimate shapley values with vision transformers. *arXiv preprint arXiv:2206.05282*, 2022.

Bernardo P de Almeida, Franziska Reiter, Michaela Pagani, and Alexander Stark. Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of synthetic enhancers. *Nature Genetics*, 54(5): 613–624, 2022.

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

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*, pp. 248–255. Ieee, 2009.

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

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

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.

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.

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

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

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

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.

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*, pp. 770–778, 2016.

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

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

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.

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

Saachi Jain, Hadi Salman, Eric Wong, Pengchuan Zhang, Vibhav Vineet, Sai Vemprala, and Aleksander Madry. Missingness bias in model debugging. *arXiv preprint arXiv:2204.08945*, 2022.

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

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pp. 2907–2916. PMLR, 2020.

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.

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.

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

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

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

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*, pp. 91–115. Springer, 2022.

Yongchan Kwon and James Zou. Weightedshap: analyzing and improving shapley based feature attributions. *arXiv preprint arXiv:2209.13429*, 2022.

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

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.

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

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.

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.

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

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

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*, pp. 23–38. PMLR, 2020.

Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pp. 17–38. Springer, 2020.

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.

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

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

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*, pp. 454–466. Springer, 2019.

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.

Guillermo Owen. *Game theory*. Emerald Group Publishing, 2013.

Raymond EAC Paley and Antoni Zygmund. On some series of functions,(3). In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pp. 190–205. Cambridge University Press, 1932.

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*, pp. 379–410, 1933.

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

Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. The shapley value in machine learning. *arXiv preprint arXiv:2202.05594*, 2022.

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.

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*, pp. 618–626, 2017.

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

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.

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

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

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

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.

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

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

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.

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

Jacopo Teneggi, Paul H Yi, and Jeremias Sulam. Weakly supervised learning significantly reduces the number of labels required for intracranial hemorrhage detection on head ct. *arXiv preprint arXiv:2211.15924*, 2022b.

Leonard Henry Caleb Tippett et al. The methods of statistics. *The Methods of Statistics.*, 1931.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Isabella Verdinelli and Larry Wasserman. Feature importance: A closer look at shapley values and loco. *arXiv preprint arXiv:2303.05981*, 2023.

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

David S Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. Explaining predictive uncertainty with information theoretic shapley values. *arXiv preprint arXiv:2306.05724*, 2023.

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

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

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

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*, pp. 2921–2929, 2016.

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.