

Challenges and Opportunities of Shapley values in a Clinical Context

Anonymous Authors¹

Abstract

With the adoption of machine learning-based solutions in routine clinical practice, the need for reliable interpretability tools has become pressing. Shapley values are a local explanations which gained popularity in recent years. Here, we reveal current misconceptions about the “true to the data” or “true to the model” trade-off and demonstrate its importance in a clinical context. We show that the interpretation of Shapley values, which strongly depends on the choice of a reference distribution for modelling feature removal, is often misunderstood. We further advocate that for applications in medicine, the reference distribution should be tailored to the underlying clinical question. Ultimately, we advise on the right reference distributions for specific medical use-cases.

1. Introduction

Since the introduction of KernelSHAP (Lundberg & Lee, 2017), Shapley values have become a standard in the model explainability literature. Originally proposed to attribute the effect that a player has on the outcome of a game, there have been some challenges in transferring this game theoretic notion to model interpretation. Given a model $f(x_1, \dots, x_m)$, the features from 1 to m can be considered players in a game in which the payoff f is a measure of the importance or influence of that subset. While Shapley values in game theory depend on the ordering of the players, Shapley values for model interpretation “remove” or drop features instead. When a feature is removed, its value is sampled from a reference distribution. When it isn’t removed, its value is set to the feature value of our instance. The choice of a reference distribution remains challenging (Aas et al., 2019; Janzing et al., 2020; Frye et al., 2020). Recent papers on Shapley values have pointed out the limitations of each type of distribution (Kumar et al., 2020; Fryer et al., 2021) causing

the Shapley values to be seen as unstable and as such, unfit for safety-critical domains such as healthcare. In contrast, we argue that Shapley values aren’t insufficient but that the main challenges arise from the misconceptions around their approach to explanation. We advocate for a careful analysis of the possible reference distributions in light of the underlying research question before using Shapley values in a medical settings.

1.1. Contributions

1. We explicitly define the clinical questions Shapley values can respond to, for each reference distribution. Contrasting with (Chen et al., 2020) who only differentiate between observational and interventional conditional reference distributions, we adapt our discussion to clinical contexts and take a broader viewpoint by integrating causality-based Shapley values.
2. We use carefully chosen counter-examples to highlight the existing misinterpretations of Shapley values and illustrate how certain choices of reference distributions may lead to misleading results. By doing so, we go beyond the misconceptions around the “true to the data” or “true to the model” trade-off. We further show that Shapley values suffer from limitations that stem from model under-performance.
3. We discuss which reference distribution is best suited for the following use-cases: counterfactual fairness, statistical fairness, explainability and feature selection; and illustrate our statements using clinical examples.

2. Background

In model explainability, Shapley values quantify the contribution of various features $\{1, \dots, m\}$ in the prediction of a complex model $f : \mathbb{R}^m \rightarrow \mathbb{R}^l$ at an instance x as follows: $f(x) = \phi_0^f(x) + \sum_{i=1}^M \phi_i^f(x)$ where $\phi_j^f(x)$ is the Shapley value of feature j to $f(x)$ and $\phi_0^f(x) = \mathbb{E}f(X)$ is the average prediction with the expectation over the observed data distribution. The contribution of a feature j is computed based on the change in value function v comparing when the feature j is equal to the value x_j with when it is removed from the input. To account for the dependence with other features, one takes the difference in value function v averaged over all possible coalitions.

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

tions S of features excluding feature j . The value function is defined as the expectation of the black box algorithm at observation x over the not-included features \bar{S} using a reference probability distribution over feature values $r(X^* | x_S)$ such that $v(S) = \mathbb{E}_{r(X^* | x_S)}[f(x_S, X_{\bar{S}}^*)]$ for $\bar{S} := \{1, \dots, m\}/S$ and the operation $(x_S, x_{\bar{S}})$ denoting the concatenation of its two arguments. Ultimately, a binomial weight $\frac{|S|!(m-|S|-1)!}{m!}$ is used to recover the original Shapley values which account for all possible orderings. As a result, the Shapley values of feature j is defined as follows:

$$\begin{aligned} \phi_j^f(x) &= \frac{1}{m} \sum_{\substack{|S| \in \\ \{0, \dots, m-1\}}} \left[\frac{1}{\binom{m-1}{|S|}} \sum_{\substack{S \subseteq \{1, \dots, m\}/j \\ \text{with } |S|=j}} \Delta v_f(S, j, x) \right] \\ &= \mathbb{E}_S \left[\mathbb{E}_{r(X^* | x_S)}[f(x_S, X_{\bar{S}}^*)] \right] \end{aligned}$$

where $\Delta v_f(S, j, x) = v_f(S \cup \{j\}, x) - v_f(S, x)$

An equivalent approach to Shapley values is to consider that the attribution of feature j equal to the expected change in the value function of including feature j after a random number of features has already been included

$$\begin{aligned} \phi_v(j) &= \frac{1}{m} \sum_{\substack{|S| \in \\ \{0, \dots, m-1\}}} \left[\frac{1}{\binom{m-1}{|S|}} \sum_{\substack{S \subseteq \{1, \dots, m\}/j \\ \text{with } |S|=j}} \phi_v(j, S) \right] \\ &= \mathbb{E}_{|S|} \left[\mathbb{E}_S [\phi_v(j, S)] \right]. \end{aligned}$$

This allows us to interpret the Shapley value of feature j as the expectation of the following experiment: (1) sample the number of included features $|S|$ uniformly, (2) sample $|S|$ features from $\{1, \dots, m\}/j$ and (3) return the change in value function when including feature j given S had been included.

Therefore, the Shapley value of feature j answers the question: “What is the expected change in the value function if we first evaluate it on a random subset of features not including j before we evaluate it additionally on j ?”. In practice, over the reference distribution is estimated using a sample mean over a finite number of reference points. Further, the choice of a reference distribution to sample from has been subject to recent debates (Aas et al., 2019; Janzing et al., 2020; Merrick & Taly, 2020). Marginal Shapley values (Lundberg & Lee, 2017; Janzing et al., 2020) define $r(X^* | x_S) := p(X^*)$ where p denotes the marginal data distribution. Conditional Shapley values (Aas et al., 2019) set the reference distribution equal to the conditional distribution given x_S $r(X^* | x_S) := p(X^* | X_S^* = x_S)$. Since Shapley values are an expectation of $v(S)$ over all coalitions. All in all, Shapley values for model interpretability are thus

defined by

$$\begin{aligned} \phi(j) &= \mathbb{E}_S \left[\mathbb{E}_{r(X^* | x_S)}[f(x_S, X_{\bar{S}}^*)] \right] \\ &= \mathbb{E}_{r(X^* | x_S)}[\phi_{x^*}(j)] \end{aligned}$$

3. Marginal, Conditional and Causality-based Shapley values

3.1. Marginal Shapley values

Lundberg et. al (2017) suggest using marginal Shapley values by sampling the dropped features \bar{S} from the data set for the computation of the value function $v(S) = \mathbb{E}_{r(X_{\bar{S}}^* | x_S)}[f(x_S, X_{\bar{S}}^*)]$. Janzing et. al (2020) show that the value function then becomes an expectation where the included features have been “intervened on”, that is, set to a value using the *do* operation $v(S) = \mathbb{E}[f(X_S, X_{\bar{S}}^*) | do(X_S = x_S)]$. In this setting, the reference distribution is $r(X_{\bar{S}}^* | x_S) = p(X_{\bar{S}}^*)$ where p is the data distribution of X . Consider a clinical model f predicting the probability of a disease from patient features. The question the marginal Shapley value of feature j for patient x answers is, “What is the expected change in predicted disease probability if we were to set feature j of a random patient x^* of the cohort equal to x_j after we already set a random subset of features (excluding j) equal to x ?”.

Marginal Shapley values are often described as “true to the model” (Chen et al., 2020), which means that attributions depend on whether the features are explicitly used by the black-box. Although this is partly true, we contrast by showing that marginal Shapley values are still heavily influenced by the distribution of the data.

Consider a linear model which predicts cardiovascular diseases $f(x) = x_1 + x_2$ where x_1 is a measure of the patient’s cholesterol relative to the recommended upper limit and x_2 is a patient’s fasting blood sugar (in g/L). We assume $x_1 \sim \text{Uniform}(-1, 2)$ and $x_2 \sim \text{Uniform}(0, 3)$ and a patient $x = (0, 0)$. The Shapley value for x_1 will be -0.5 and the Shapley value for x_2 will be -1.5 (see Supplementary Material A.1). Here, the difference in magnitude between the two feature attributions results solely from the fact that they are differently far from their mean. In other words, it stems from the fact that the outcome of the model for our patient is differently far from the expected outcome when intervening on cholesterol compared to when intervening on fasting blood sugar value. Further, note that for the same black box model f if x_1 and x_2 are centred on a common mean but have different spreads their two marginal Shapley values would not be equal, even though they play symmetric roles in the algebraic formulation of the model (see Supplementary Material A.2). Ultimately, for similar reasons if away from the center of mass, marginal Shapley values may lack locality (Ghalebikesabi et al., 2021).

Therefore, marginal Shapley values may still acknowledge the overall distribution of features and where our observation stands in this distribution, instead of depending solely on the model and the patient’s features. In other words, the explanation given would be *relative* to the rest of the cohort, and not faithful to the model only. As model faithfulness is paramount in certain clinical contexts e.g. when evaluating the safety of a model, using marginal Shapley in these scenarios raises ethical concerns (Merrick & Taly, 2020).

3.2. Conditional Shapley values

Conditional Shapley values have value function $v(S) = \mathbb{E}_{p(X_S^*|x_S)}[f(X_S, X_S^*) \mid X_S = x_S]$ where the reference function is the conditional distribution $p(X_S^*|x_S)$. For a clinical predictive model, these values answer the question, “What is the expected difference in model outcome between a cohort with the same characteristics as our patient x for a random subset of features including j and a cohort that has the same characteristics as our patient in the same random subset of features excluding j ?”. Alternatively, (Covert et al., 2020) show that using a conditional reference distribution can also be argued for from an information theoretic viewpoint: the model outcome of certain risk minimising black box models retrained on a restricted feature set can be approximated by the conditional expectation of the complete model outcome (Covert et al., 2020). From this view, conditional Shapley values answer the question, “What is the expected change in a model’s prediction assuming it was trained on a restricted feature set (excluding j) and we retrained it on the restricted feature set including j ?”. Conditional Shapley values are often considered to be “true to the data”, meaning that the model isn’t evaluated on out-of-distribution instances i.e. implausible synthetic patients. Conditioning on the present features prevents from breaking the feature correlations observed in the reference population (Chen et al., 2020; Frye et al., 2020; Covert et al., 2020).

3.3. Causality-based Shapley Values

Interest for causal explanations has increased in the scientific community. Consequently, several approaches to causality based computations of Shapley values have been introduced in recent years. We refer to these as causality-based Shapley values. Such methods only consider causally-consistent orderings of single-reference games i.e. orderings such that known causal ancestors precede their descendants. End users are thus able to compute feature attributions even when they have partial knowledge of the causal graph underlying the data. These types of approaches favour explanations in terms of root causes, as they quantify the impact a given feature has on model prediction while its descendants remain unspecified. Asymmetric Shapley (Frye et al., 2019) uses conditioning by observation whereas Causal

Shapley (Heskes et al., 2020) uses conditioning by intervention. Asymmetric Shapley values are said to be better for dataset with intrinsic ordering, such as temporal data, but Causal Shapley values are less sensitive to misspecifications of the causal graph structure (Heskes et al., 2020). We notice that causality-based Shapley values strongly rely on the assumed graph. If the true causal structure deviates, the resulting explanation may be false. As such, causality-based Shapley values may “force” explanations to fit to the hypothesised causal flow.

4. Use cases for reference distributions

4.1. Counterfactual fairness

One of the main benefits of marginal Shapley values is their dummy property, which says that only features that are in the model will get a non-zero attribution (Janzing et al., 2020; Sundararajan & Najmi, 2020). This property is of interest when examining the fairness of an algorithm. In particular, the notion of counterfactual fairness requires that for any individual, their model outcome remains unchanged if the *sensitive* attribute (e.g. race, gender) had another value (Kusner et al., 2017). Counterfactual fairness is essential for any clinical application as using opaque decision making models may introduce or exacerbate disparities in care (Ahmad et al., 2020). Inspired by the example used from Merrick et. al (2020), we consider f to be a black box algorithm for clinical management of a disease $f(x) = x_{male}$ with two binary features x_{male} and $x_{eligible}$ where the latter indicates that a patient is eligible to surgery i.e. is likely to benefit from it. We further assume males are more likely to be eligible to surgery. Since intervening on the variable $x_{eligible}$ and setting it to any arbitrary value will not have any direct effect on the model outcome, $x_{eligible}$ receives a marginal Shapley value of 0. By contrast, x_{male} receives a marginal value of 1, as it is equal to the expected difference of the model outcome of the male population (here 1) and a population that has the same value as x either in $x_{eligible}$ or not in any feature.

Proposition *Let f be a black box model and feature i be a sensitive attribute, having $\phi_i=0$, where ϕ_i denotes the marginal Shapley value of feature i , is a necessary, non sufficient condition to f being a counterfactually fair w.r.t feature i .*

The proof of the above result is in the Supplementary Material B, where we use an unfair function $f(x) = x_{male}x_{eligible}$ in an all-male population and show the marginal attribution for gender is trivially 0. On the other hand, conditional Shapley values cannot be used to determine counterfactual fairness. In the previous examples, the conditional attribution of $x_{eligible}$ is equal to a weighted sum of (a) the expected difference between

the model outcome of the population which is eligible and the common population, and (b) the expected difference in model outcome between the population which is male and is eligible and the population which is male. Term (b) is null as both terms equal 1. Since the eligible sub-population will more likely be male than the common population, term (a) will be positive. Therefore, using conditional Shapley renders a positive attribution for $x_{eligible}$ and the model will seem fairer than it is.

4.2. Statistical fairness

Statistical fairness covers criteria where we are interested in the expected difference in model outcome between the population with a specific sensitive attribute and the general population. These types of criteria – such as having equal positive rates – are popular in current medical AI practice as they tend to be easier to evaluate (Hedden, 2021). Frye et. al (2020) show that unfair models could be rephrased on unprotected features that are correlated with a sensitive attribute such that the outcome of the model remains the same for observations on the data manifold. Two models with different algebraic formulations but which evaluate the same on the data can thus render highly different marginal Shapley values whilst keeping close conditional Shapley values. This makes it challenging to judge the statistical fairness of an algorithm using a framework based on feature removal, especially in sensitive environments.

4.3. Feature selection

Feature selection is essential to build parsimonious models for clinicians to work with, particularly in resource-limited contexts. For instance, a risk score is a measure of clinical severity. It is often built directly from the feature attributions computed based on a prognosis model (Olson et al., 2013; Bernabe-Ortiz et al., 2016). Here, explanation models are used for selecting the most important variables in order to obtain a simplistic scoring system.

The information theoretic equivalence regarding conditional Shapley values shown by Covert et. al (2020) entails implies these can be used for feature selection, as a single coalition Shapley value of 0 for feature j reflects that there is no change in model outcome from including feature j in the training process. We disagree with that interpretation, as having a Shapley value of 0 is still possible when including j has a small effect in some coalitions, and a greater impact in others. Consider for instance the Shapley value of feature 2 in $f(x) = \mathbb{I}(x_1 > 1) \cdot 3x_2 - \mathbb{I}(x_1 \leq 1) \cdot x_2$ at $x = (0.5, 0.5)$. Including feature 2 decreases the model outcome as we know that $x_1 = 0.5$ but in expectation over feature 1, including feature 2 will increase the model outcome (see Supplementary Material C for a proof). Further, Covert et. al (2020) note that their interpretation only holds for a

restricted set of black box functions. Assuming we only have black box access to the function, we would not know if the model under study exhibits the desired properties. Ultimately, the computational burden of Shapley values for feature selection is heavy, as it requires computing all instance-specific Shapley values in the global population.

4.4. Justification

Explainability can refer to a variety of concepts (Adadi & Berrada, 2018). In the following, we focus on situations where one (e.g. a clinician) needs to *justify* a decision. For such a purpose, using marginal Shapley values may be misleading. In the former clinical management example if the physician does not have access to the eligibility of a patient but to their weight, the latter might be used as a proxy measure of their eligibility to the surgery and have non-zero feature attribution. A common misconception about conditional Shapley values is that they would correctly assign a non-zero attribution to strength due to the fact that they are “true to the data” (Chen et al., 2020), contrasting with their marginal equivalent. We claim that such a description is reductive as it overlooks the fact that Shapley values explain the expected change in *model* outcome. In other words, if the black box model does not capture the data dependence between the target variable and the covariates, for instance, because of mis-specification or low performance, even conditional Shapley values will not be true to the data.

Ultimately, a clinical decision should be explained in a *causal* manner (Tonekaboni et al., 2019). Introducing causal knowledge into the explanation model should prevent spurious associations from being given importance. Therefore, we argue that for justifying a decision (Adadi & Berrada, 2018), causality-based Shapley values (Frye et al., 2019; Heskes et al., 2020; Wang et al., 2021) should be used. However, existing methods strongly rely on causal assumptions. We thus caution against using them when domain knowledge is insufficient to build a causal graph, and further add that the black box to be explained should have a high predictive performance. Ultimately, we claim that further work is needed in this space for providing robust causal interpretations.

5. Concluding remarks

We give an overview of the existing misconceptions around Shapley values and explain how they can translate into potentially harmful explanations in a clinical context. We further define use cases for each reference distributions. To the best of our knowledge, this is the first attempt at discussing the right explanation technique for a targeted clinical question. Ultimately, we claim that further work should be done on explainable methods for models that are *causal* themselves.

References

- Aas, K., Jullum, M., and Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *arXiv preprint arXiv:1903.10464*, 2019.
- Adadi, A. and Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.
- Ahmad, M. A., Patel, A., Eckert, C., Kumar, V., and Tere-desai, A. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3529–3530, 2020.
- Bernabe-Ortiz, A., Smeeth, L., Gilman, R. H., Sanchez-Abanto, J. R., Checkley, W., Miranda, J. J., et al. Development and validation of a simple risk score for undiagnosed type 2 diabetes in a resource-constrained setting. *Journal of diabetes research*, 2016, 2016.
- Chen, H., Janizek, J. D., Lundberg, S., and Lee, S.-I. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.
- Covert, I., Lundberg, S., and Lee, S.-I. Explaining by removing: A unified framework for model explanation. *arXiv preprint arXiv:2011.14878*, 2020.
- Frye, C., Feige, I., and Rowat, C. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*, 2019.
- Frye, C., de Mijolla, D., Cowton, L., Stanley, M., and Feige, I. Shapley-based explainability on the data manifold. *arXiv preprint arXiv:2006.01272*, 2020.
- Fryer, D., Strümke, I., and Nguyen, H. Shapley values for feature selection: The good, the bad, and the axioms. *arXiv preprint arXiv:2102.10936*, 2021.
- Ghalebikesabi, S., Ter-Minassian, L., DiazOrdaz, K., and Holmes, C. C. On locality of local explanation models. *Advances in Neural Information Processing Systems*, 34, 2021.
- Hedden, B. On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, 49(2), 2021.
- Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *arXiv preprint arXiv:2011.01625*, 2020.
- Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916. PMLR, 2020.
- Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pp. 5491–5500. PMLR, 2020.
- Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.
- Lundberg, S. and Lee, S.-I. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.
- Merrick, L. and Taly, A. The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 17–38. Springer, 2020.
- Olson, D., Davis, N. L., Milazi, R., Lufesi, N., Miller, W. C., Preidis, G. A., Hosseinipour, M. C., and McCollum, E. D. Development of a severity of illness scoring system (inpatient triage, assessment and treatment) for resource-constrained hospitals in developing countries. *Tropical Medicine & International Health*, 18(7):871–878, 2013.
- Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *International Conference on Machine Learning*, pp. 9269–9278. PMLR, 2020.
- Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pp. 359–380. PMLR, 2019.
- Wang, J., Wiens, J., and Lundberg, S. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 721–729. PMLR, 2021.