Towards Cotenable and Causal Shapley Feature Explanations

Tony Liu,[1] **Lyle Ungar**[1]
[1]Department of Computer Science, University of Pennsylvania
liutony@seas.upenn.edu, ungar@cis.upenn.edu

Abstract

A key component of being able to trust machine learning models in medical contexts is the ability to explain why the model is making a particular prediction. Feature importance methods based on Shapley values have become popular, but there has also been recent debate surrounding whether their mathematical properties may limit their use for expaining models. We outline some properties that a model explanation needs in order to be useful: model explanations should be *causal*, as ultimately they are used to aid in decision making, and they should be *cotenable*: they should respect the observed correlation between features. We show how different implementations of Shapley-based feature importances trade off these properties and propose using medical domain knowledge to group features as a step towards satisfying both causality and cotenability, which would provide model explanations that are more useful in clinical settings.

Motivation

As complex machine learning models continue to be developed for high-stakes clinical settings (Litjens et al. 2017; Sendak et al. 2020), being able to explain model output is critical. Feature importance measures based on Shapley values from cooperative game theory (Shapley 1953) show promise because they can provide post-hoc interpretation of black-box models and also satisfy desirable theoretical axioms (Lundberg and Lee 2017). However, there are variations on how to calculate Shapley-based feature importances in practice, and previous works have explored mathematical problems that arise with particular implementations which could result in misleading interpretations (Kumar et al. 2020b; Sundararajan and Najmi 2020). Shapley-based explanation implementations need to be carefully considered before being potentially used in clinical settings.

What makes a feature importance measure, such as those calculated through Shapley values, a useful model explanation for medical decision support? We highlight two properties we believe to be important, using a running example of a machine learning model trained to predict patient mortality. First, we want our measures to capture the **causal effect** of a feature on model output: if we *intervene* on a patient's cholesterol and hold their other characteristics constant, how does that change the model's prediction of the patient's mortality rate? Second, we want our measures to respect dependencies between features seen in the real world: if BMI, height, and weight are all included in our mortality model, feature importance measures of BMI that break its correlation with height and weight lose their meaning, because it is not possible to change BMI without changing one of height or weight. We refer to feature importance measures that respect feature dependencies as satisfying **cotenability**, from the philosophy of conditional logic (Arlo-Costa 2007). Note that causal explanations violate cotenability if the model is trained on correlated features, since one can query the model about changes that violate the correlation structure of the training data, and so these two properties will often have to be traded off.

Shapley value feature importances reflect this trade-off between causality and cotenability, which has also been framed as the explanation being "true to the model" vs. "true to the data" (Chen et al. 2020). When calculating Shapley values for a given feature in practice, one must decide whether to use a marginal or conditional distribution to sample the other features – sampling values from a marginal distribution satisfies causality, while sampling from a conditional distribution satisfies cotenability. We will refer any Shapley-based method using the conditional distribution as a *conditional Shapley method* and any method using the marginal distribution as a *interventional Shapley method.*

Here we review the strengths and weaknesses of both Shapley formulations and the problems they pose through a graphical interpretation. We then propose grouping features through domain expertise before generating Shapley value explanations to alleviate some problems present in interventional Shapley methods, making a step towards satisfying both cotenability and causality.

Shapley value background

We first briefly describe Shapley value calculations for feature importance. Formalized notation and theoretical axioms can be found in (Lundberg and Lee 2017; Sundararajan and Najmi 2020). Given a collection of N features, a model f, the Shapley value calculation assigns an importance value to a feature i by asking how much that feature contributes to the model output f in the presence of other features. Specifically, given an example x and a *value function* $v_{f,x}(S)$ that

maps subsets of features S to a real value, the Shapley value of a feature i is the normalized sum of its *marginal contribution* over all subsets of features S:

$$v_{f,x}(S \cup \{i\}) - v_{f,x}(S) \tag{1}$$

In order to compute Shapley values for feature attribution, we need to define the value function $v_{f,x}(S)$, and the specific choice of $v_{f,x}(S)$ produces many variations of Shapley feature importance implementations, including both interventional and conditional Shapley methods. We define the following notation: let S be the set of features we are interested in, X_S the collection of random variables associated with features in S, x_S a particular setting of the variables in X_S, and $B = N \setminus S$.

Cotenability and causality through graphs

We now give a graphical interpretation (Pearl 2009) of interventional and conditional Shapley value functions following the arguments presented in (Janzing, Minorics, and Bloebaum 2020; Zhao and Hastie 2019), which clarify how they tie to the concepts of cotenability and causality. We will consider a model f that takes two features as input $N = \{X_1, X_2\}$, representing the model output as $Y = f(X_1, X_2)$. In addition, we also consider a latent variable Z that is a common parent of all the input features X_1, X_2 (Figure 1). In the following examples, we are interested in the value function for $S = \{X_1\}$.

Conditional Shapley methods The conditional Shapley value function is defined as:

$$v_{f,x}(S) = E_{X_B|X_S}[f(X_N)|X_S = x_S] \tag{2}$$

where the expectation of the model output f is taken over a conditional distribution of X_B given a particular setting x_s of X_S from x. Shapley-based feature importance measures that use this value function include (Frye, Feige, and Rowat 2019; Aas, Jullum, and Løland 2020).

We show the graphical representation of the conditional Shapley value function for $S = \{X_1\}$ in Figure 1a. Note that the distribution of X_2 changes given the setting of $X_1 = x_1$ (Eq 2) because of their shared dependence Z.

Interventional Shapley methods The interventional Shapley value function is defined as:

$$v_{f,x}(S) = E_{X_B}[f(x_S; X_B)] \tag{3}$$

where the expectation of the model output f is taken over a marginal distribution of X_B, setting the variables in S to their corresponding values x_S. Shapley-based feature importance measures that use this value function include (Datta, Sen, and Zick 2016; Merrick and Taly 2020).

Because interventional Shapley methods compute the expected value of f over the marginal distribution of B, the resulting state of the graph in our example is the *interventional* distribution over the features (Figure 1b), as we are breaking the dependence between X_1 and X_2 by forcing X_1 to take on the value x_1. We can use the backdoor criterion (Pearl 2009) to show that $v_{f,x}(\{X_1\})$ under interventional Shapley is equivalent to the causal quantity $E[Y|do(X_1 = x_1)]$.

Satisfying causality Interventional Shapley methods captures the causal effect of setting $X_1 = x_1$ on the model output Y, which is an important property for a feature explanation: we want to know how forcing a feature to be a certain value affects the model output, not simply observe the model output when the feature happens to be that value. This is not the case for conditional Shapley methods, as they may assign importance to irrelevant features (ones that do not affect the model output) if the irrelevant features are correlated with other features used by the model (Sundararajan and Najmi 2020; Janzing, Minorics, and Bloebaum 2020).

For example, suppose the mortality model we want to explain is given both diastolic (X_1) and systolic (X_2) blood pressure as features, but the model only uses systolic blood pressure to make predictions Y. In our graphical representation, this would mean that there is no edge from X_1 to Y. Conditional Shapley methods will still attribute non-zero feature importance to diastolic blood pressure because the blood pressure measurements are correlated, which would be misleading. In medical contexts, maintaining a causal interpretation of features is critical, as machine learning models are ultimately used for decision making: what happens to patients' mortality rate if we *change* their blood pressure?

Satisfying cotenability In the graphical framework we have presented, a cotenable explanation respects the dependencies of the model inputs (the X's) on their parents (Z) by preserving all outgoing edges from Z. We can think of Z as a property of the real world that governs the relationship between X_1 and X_2, such as some characteristic of an individual's cardiovascular health (e.g. family history) that influences both systolic and diastolic blood pressure. Because intervening on a feature breaks any shared dependencies, interventional Shapley methods do not satisfy cotenability, while conditional Shapley methods do.

Failing to satisfy cotenability presents two issues in practice. First, because machine learning models learn the correlation structure among features from their training data, breaking any dependencies results in asking the machine learning model to make predictions on samples outside of the data distribution it was trained on. which may result in the model behaving erratically (Kumar et al. 2020b; Hooker and Mentch 2019).

Second, feature importances that are not cotenable may be limited in their interpretability. Returning to the situation where diastolic and systolic blood pressure are our features X_1 and X_2, suppose we use an interventional Shapley method to compute feature importances. Interventions made to blood pressure in the real world such as the prescription of diuretics often do not specifically target systolic or diastolic blood pressure in isolation, so knowing how a patients' mortality changes when changing their diastolic blood pressure while holding their systolic blood pressure constant (or vice versa) is less useful for medical decision makers.

Grouping features for explanation

We see that the choice of using conditional or interventional Shapley methods trade off the properties of cotenability and causality, yet both properties are critical components of a

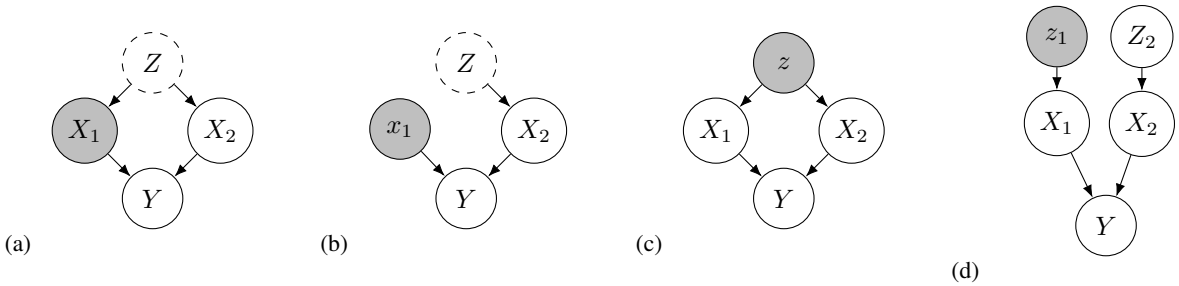

Figure 1: **Graphical representation of model input/output, causality, and cotenability.** The dotted nodes represent latent variables that are not observable. Solid border nodes represent variables that we do observe, and shaded in nodes represent particular settings of variables.

safe and effective feature importance explanation in medical contexts. Ideally, if we could intervene on the upstream variable Z, $E[Y|do(Z = z)]$ (Figure 1c), we would satisfy both cotenability and causality. Knowledge of the relationship between the parent variable Z and the model inputs X must be provided in order to perform this cotenable intervention, indicated by the solid, shaded in node for the setting $Z = z$.

An assumption of feature independence is often made when interpreting Shapley values (Janzing, Minorics, and Bloebaum 2020; Lundberg and Lee 2017), which alleviates the cotenability issue of interventional Shapley methods by treating each model input X_i as having its own upstream "cause" Z_i (Figure 1d). However, the assumption of feature independence is often unreasonable in practice. We thus argue that an important step in utilizing Shapley values is to *group* features such that the groupings are independent of one another. In the ideal case where groups are completely independent, this will effectively produce the graph depicted in Figure 1d, satisfying both cotenability and causality. These groupings should be informed by domain expertise, which can reflect causal knowledge of data as well as dependencies among features the practitioner may wish to impose. For example, instead of considering systolic and diastolic blood pressure separately we may group them together before calculating Shapley value feature importances. By grouping features, we may mitigate instability in model output due to violations of cotenability.

Grouping features may be useful beyond respecting cotenability, as they can increase the interpretability of the feature explanations provided. In practice, it is often difficult to precisely intervene on one feature input into the model and grouping variables together that are more easily intervened on in aggregate (such as our general "blood pressure" grouping) enables more actionable explanations.

Case study: NHANES I survival prediction

We now perform a simple case study using the NHANES I data (CDC 1974) to predict survival that illustrates the interpretability benefits of grouping features.

Qualitative feature groupings To understand the relationship between features present in the NHANES I dataset, we first compute Pearson correlations between all of the predictors, and then perform a hierarchical clustering on the resulting correlations. We use domain knowledge provided by a medical student to create qualitative groupings of the features (Figure 2). Though features related to blood pressure ("blood pressure") are nested within the "cardiovascular risk" block, we break them out as a separate group as an illustration of feature cotenability: pulse pressure is the difference between systolic and diastolic blood pressure.

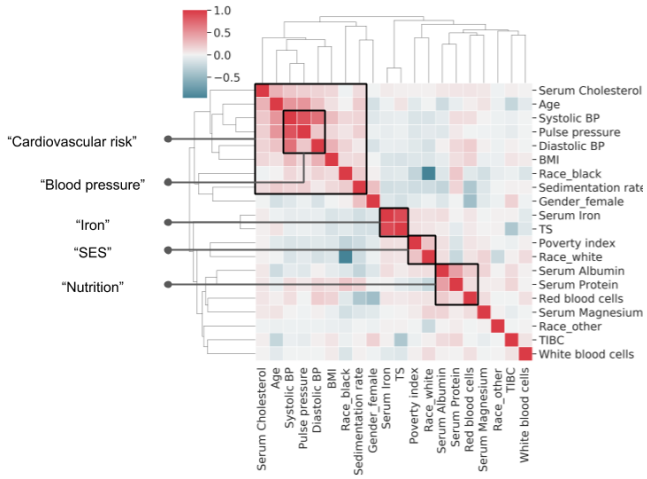

Figure 2: **Hierarchical clustering of NHANES I features with qualitative groupings**. Cells in the clustermap are colored based on their Pearson correlation.

Model and feature setup Our feature groupings for model input are "cardiovascular risk" without blood pressure features, "blood pressure," "iron," "SES" (socioeconomic status), and "nutrition," corresponding to the groups in Figure 2, which are combined as a standardized sum of features within each group. We train random forest models and use an interventional Shapley method (Lundberg et al. 2020) to calculate feature importances.

Grouping improves importance interpretability Shapley feature importances computed on individual features pose an additional problem in practice – importance values may be spread across correlated features, as shown by the Shapley importances for the individual systolic BP, diastolic

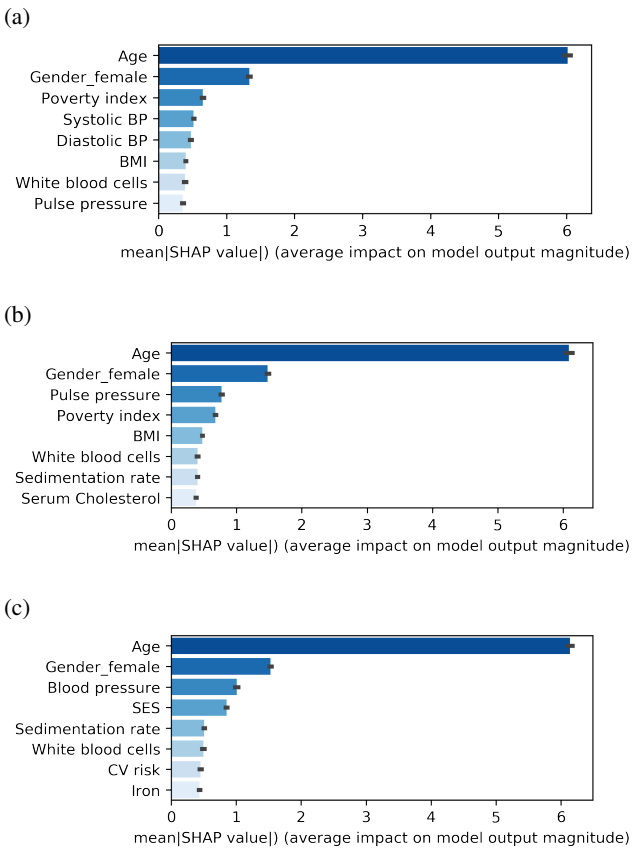

Figure 3: **Interventional Shapley feature importances when including (a) all blood pressure features individually, (b) only pulse pressure, and (c) blood pressure as a grouped feature.** Note that blood pressure appears less important in (a). Error bars are generated over ten sub-samples of held-out testing data.

BP, and pulse pressure features (Figure 3a). By considering these features individually, their relative importance is diluted, all ranking below poverty index. Removing correlated features alters the relative importance rankings of features in the model: by including pulse pressure as the only blood pressure feature in the training data, we see it becoming the third most important feature in predicting survival (Figure 3b), despite it being ranked much lower when the other blood pressure features were included. Grouping all the blood pressure measurements into a single feature for model input similarly results in the relative importance of "blood pressure" being higher than all the features other than age and gender (Figure 3c). Grouping correlated features can aid in determining relative importance of model inputs based on their Shapley value.

Additionally, the grouped features can help produce more actionable feature importance interpretability. As previously discussed, the blood pressure measures should be considered together not only because of their observed correlation structure but also because they map onto actions that can be taken by healthcare providers: interventions on patient often target blood pressure generally, not their specific systolic, diastolic, or pulse pressure values. The groupings overall can transport more information to a wider audience. For example, "cardiovascular risk," which consists of BMI, serum cholesterol, and race, can reduce cognitive load and be a more meaningful label for physicians to use when examining a mortality model. Combining features into groups that respect cotenability increases overall interpretability of model explanations.

Discussion

Here we have reviewed conditional and interventional Shapley methods and show how they trade off desirable properties of model explanation through a graphical lens. Conditional Shapley methods do not provide a causal interpretation of the feature importances, while interventional Shapley methods violate cotenability among the features. We proposed grouping features under interventional Shapley as step towards satisfying both properties. As the features should be grouped through domain knowledge in order to produce cotenable and actionable interventions, this necessitates integrating end user medical professionals into the machine learning model building process.

Other limitations of Shapley-based feature importance Though we highlight cotenability and causality as necessary components for a useful Shapley-based explanation, there are other issues with their implementation that we have not covered. In practice, we usually cannot evaluate the expectations presented in Equations 2 and 3, and so they are often approximated by some empirical distribution of the training data. The choice of how to sample from these empirical distributions raise additional questions about interpretability and stability of Shapley-based importances (Sundararajan and Najmi 2020). Furthermore, (Kumar et al. 2020b) find limitations in Shapley-value explanations for nonlinear models, due to additivity constraints of Shapley value calculations. These issues must also be resolved in order to apply Shapley-based feature importances safely in practice.

Quantifying Shapley explanation quality In addition to examining the qualitative benefits of increased interpretation through grouping features, we also would like to quantify the quality of a Shapley explanation. Shapley residuals (Kumar et al. 2020a) are a promising metric that can measure the extent of interventional effects of changing a feature on model output. Future work will explore using Shapley residuals to measure the quality of grouped Shapley features – feature groupings that better respect cotenability are "more independent" from one another, which could be captured by a lower Shapley residual.

Conclusion Though cotenability and causality are both important properties of a useful model explanation, we see that the choice of Shapley method necessitates a trade-off between them. We believe that work on grouped Shapley values can bridge the gap and move towards satisfying both properties, enabling more actionable Shapley-based feature explanations in medical settings.

References

Aas, K.; Jullum, M.; and Løland, A. 2020. Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values. *arXiv:1903.10464 [cs, stat]* .

Arlo-Costa, H. 2007. The logic of conditionals. https://plato.stanford.edu/entries/logic-conditionals/.

CDC. 1974. NHANES I. National Health and Nutrition Examination Survey. https://wwwn.cdc.gov/nchs/nhanes/nhanes1/Default.aspx.

Chen, H.; Janizek, J. D.; Lundberg, S.; and Lee, S.-I. 2020. True to the Model or True to the Data? .

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. ISSN 2375-1207. doi:10.1109/SP.2016.42.

Frye, C.; Feige, I.; and Rowat, C. 2019. Asymmetric Shapley Values: Incorporating Causal Knowledge into Model-Agnostic Explainability. *arXiv:1910.06358 [cs, stat]* .

Hooker, G.; and Mentch, L. 2019. Please Stop Permuting Features: An Explanation and Alternatives. *arXiv:1905.03151 [cs, stat]* URL http://arxiv.org/abs/1905.03151. ArXiv: 1905.03151.

Janzing, D.; Minorics, L.; and Bloebaum, P. 2020. Feature Relevance Quantification in Explainable AI: A Causal Problem. In *International Conference on Artificial Intelligence and Statistics*, 2907–2916. ISSN 2640-3498.

Kumar, I. E.; Scheidegger, C.; Venkatasubramanian, S.; and Friedler, S. A. 2020a. Shapley Residuals: Quantifying the limits of the Shapley value for explanations 9.

Kumar, I. E.; Venkatasubramanian, S.; Scheidegger, C.; and Friedler, S. 2020b. Problems with Shapley-Value-Based Explanations as Feature Importance Measures. *arXiv:2002.11097 [cs, stat]* .

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J. A. W. M.; van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42: 60–88. ISSN 1361-8415. doi:10.1016/j.media.2017.07.005. URL http://www.sciencedirect.com/science/article/pii/S1361841517301135.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2(1): 56–67. ISSN 2522-5839. doi:10.1038/s42256-019-0138-9. URL http://www.nature.com/articles/s42256-019-0138-9. Number: 1 Publisher: Nature Publishing Group.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 4765–4774. Curran Associates, Inc. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Merrick, L.; and Taly, A. 2020. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. *arXiv:1909.08128 [cs, stat]* .

Pearl, J. 2009. *Causality*. Cambridge university press.

Sendak, M. P.; Ratliff, W.; Sarro, D.; Alderton, E.; Futoma, J.; Gao, M.; Nichols, M.; Revoir, M.; Yashar, F.; Miller, C.; Kester, K.; Sandhu, S.; Corey, K.; Brajer, N.; Tan, C.; Lin, A.; Brown, T.; Engelbosch, S.; Anstrom, K.; Elish, M. C.; Heller, K.; Donohoe, R.; Theiling, J.; Poon, E.; Balu, S.; Bedoya, A.; and O'Brien, C. 2020. Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. *JMIR Medical Informatics* 8(7): e15182. doi:10.2196/15182. URL https://medinform.jmir.org/2020/7/e15182/. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

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

Sundararajan, M.; and Najmi, A. 2020. The Many Shapley Values for Model Explanation. *arXiv:1908.08474 [cs, econ]* .

Zhao, Q.; and Hastie, T. 2019. Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics* 0(0): 1–10. ISSN 0735-0015. doi:10.1080/07350015.2019.1624293.