SUFFICIENT AND NECESSARY EXPLANATIONS (AND WHAT LIES IN BETWEEN)

Anonymous authors
Paper under double-blind review

ABSTRACT

As complex machine learning models continue to be used in high-stakes decision settings, explaining their predictions is crucial. Post-hoc explanation methods aim to identify which features of an input $\mathbf{x}$ are important to a model's prediction $f(\mathbf{x})$. However, explanations often vary between methods and lack clarity, limiting the information we can draw from them. To address this, we formalize two precise concepts—*sufficiency* and *necessity*—to quantify how features contribute to a model's prediction. We demonstrate that, although intuitive and simple, these two types of explanations may fail to fully reveal which features a model considers important. To overcome this, we propose and study a unified notion of importance that spans the entire necessity-sufficiency axis. Our unified notion, we show, has strong ties to other popular notions of feature importance, like those based on conditional independence and game-theoretic quantities like Shapley values. Lastly, through various experiments, we demonstrate that generating explanations along the necessity-sufficiency axis can uncover important features that may otherwise be missed and reveal that many post-hoc methods only provide features that are sufficient rather than necessary.

1 INTRODUCTION

Over recent years, modern machine learning (ML) models, mostly deep learning-based, have achieved impressive results across several complex domains. Models can now solve difficult image classification, inpainting, and segmentation problems, perform accurate text and sentiment analysis, predict the three-dimensional conformation of proteins, and more (LeCun et al., 2015; Wang et al., 2023). Despite their success, the rapid integration of these models into society requires caution (The White House, 2023). Modern ML systems are black-boxes, comprised of millions of parameters and non-linearities that obscure their prediction-making mechanisms from everyone. This lack of clarity raises concerns about explainability, transparency, and accountability (Zednik, 2021; Tomsett et al., 2018). Thus, understanding how these models work is essential for their safe deployment.

The lack of explainability has spurred research efforts in eXplainable AI (XAI), with a major focus on developing post-hoc methods to explain black-box model predictions, especially at a *local* level. For a model f and input $\mathbf{x} \in \mathbb{R}^d$, these methods aim to identify which features in $\mathbf{x}$ are *important* for the model's prediction, $f(\mathbf{x})$. They do so by estimating a notion of importance for each feature (or groups), which allows for a ranking of importance. Popular methods include CAM (Zhou et al., 2016), LIME (Ribeiro et al., 2016), gradient-based approaches (Selvaraju et al., 2017; Shrikumar et al., 2017; Jiang et al., 2021), rate-distortion techniques (Kolek et al., 2022), Shapley value-based explanations (Chen et al., 2018b; Teneggi et al., 2022; Mosca et al., 2022), perturbation-based methods (Fong & Vedaldi, 2017; Fong et al., 2019; Dabkowski & Gal, 2017), among others (Chen et al., 2018a; Yoon et al., 2018; Jethani et al., 2021; Wang et al., 2021; Ribeiro et al., 2018). However, many of these approaches lack rigor, as the meaning of their computed scores is often ambiguous. For example, it's not always clear what large or negative gradients signify or what high Shapley values reveal about feature importance. To address these concerns, other research has focused on developing explanation methods based on logic-based definitions (Ignatiev et al., 2020; Darwiche & Hirth, 2020; Darwiche & Ji, 2022; Shih et al., 2018), conditional hypothesis testing Teneggi et al. (2023); Tansey et al. (2022), among formal notions. While these methods are a step towards rigor, they have drawbacks, including reliance on complex automated reasoners and limited ability to communicate their results in an understandable way for human decision-makers.

In this work, we advance XAI research by providing formal mathematical definitions of *sufficient* and *necessary* features for explaining complex ML models. First, we illustrate how, although informative, sufficient and necessary explanations offer incomplete insights into feature importance. To address this, we propose and study a more general unified framework for explaining models. Finally, we offer two novel perspectives on our framework through the lens of conditional independence and Shapley values, and crucially, show how it reveals new insights into feature importance.

1.1 Summary of our Contributions

We propose and study two approaches, sufficiency, and necessity, which evaluate the contribution of a set of features in $\mathbf{x}$ toward a model prediction $f(\mathbf{x})$. A sufficient set preserves the model's output, while a necessary set, when removed, renders the output uninformative. Although the two concepts appear complementary, their precise relationship remains unclear. How similar are sufficient and necessary subsets? How different? To address these questions, we study the two concepts and propose a *unification* of both. Our contributions are summarized as follows:

1. We formalize precise mathematical definitions of sufficient and necessary features for model predictions that are related but complementary to those in previous works.
2. We propose a unified approach that combines sufficiency and necessity, exploring when and how they align or differ. Additionally, we motivate its utility by highlighting its connections to conditional independence and Shapley values, a game-theoretic measure of feature importance.
3. Through experiments of increasing complexity, we demonstrate how a unified perspective uncovers new, significant, and more comprehensive insights into feature importance.

2 Sufficiency and Necessity

Notation & Setting. We use boldface uppercase letters to denote random vectors (e.g., $\mathbf{X}$) and lowercase for their values (e.g., $\mathbf{x}$). For a subset $S \subseteq [d] := \{1, \ldots, d\}$, we denote its cardinality by $|S|$ and its complement $S^c = [d] \setminus S$. Subscripts index features; e.g., $\mathbf{x}_S$ represents $\mathbf{x}$ restricted to the entries indexed by S. We consider a supervised learning setting with an unknown distribution $\mathcal{D}$ over features $\mathcal{X} \subseteq \mathbb{R}^d$ and labels $\mathcal{Y} \subseteq \mathbb{R}$. We assume access to a model $f : \mathcal{X} \mapsto \mathcal{Y}$ that was trained on samples from $\mathcal{D}$. For an input $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, the goal is to identify the important features in $\mathbf{x}$ for the prediction $f(\mathbf{x})$. To define importance, we will use the average restricted prediction, $f_S(\mathbf{x}) = \underset{\mathbf{X}_{S^c} \sim \mathcal{V}_{S^c}}{\mathbb{E}} [f(\mathbf{x}_S, \mathbf{X}_{S^c})]$, where $\mathbf{x}_S$ is fixed and $\mathbf{X}_{S^c}$ is a random vector drawn from an arbitrary reference distribution $\mathcal{V}_{S^c}$ (which may or may not depend on S^c). For example, two common choices are the marginal $\mathcal{V}_{S^c} = p(\mathbf{X}_{S^c})$ and conditional distribution $\mathcal{V}_{S^c} = p(\mathbf{X}_{S^c} \mid \mathbf{x}_S)$. This strategy, popularized in (Lundberg & Lee, 2017; Lundberg et al., 2020), allows us to query f, which only takes inputs in $\mathbb{R}^d$, and analyze its behavior when sets of features are retained or removed.

Definitions. We now present our proposed definitions of sufficiency and necessity. At a high level, these definitions were formalized to align with the following guiding principles:

P1. S is sufficient if it is enough to generate the original prediction, i.e. $f_S(\mathbf{x}) \approx f(\mathbf{x})$.

P2. S is necessary if we cannot generate the original prediction without it, i.e. $f_{S^c}(\mathbf{x}) \not\approx f(\mathbf{x})$.

P3. The set $S = [d]$ should be maximally sufficient and necessary for $f(\mathbf{x})$.

The principles P1 and P2 are natural and agree with the logical notions of sufficiency and necessity. Furthermore, because the full set of features provides all the information needed to make the prediction $f(\mathbf{x})$, it should thus be regarded as maximally sufficient and necessary (P3). With these principles laid out, we now formally define sufficiency and necessity.

Definition 2.1 (Sufficiency). *Let $\epsilon \geq 0$ and let $\rho : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ be a metric on $\mathbb{R}$. A subset $S \subseteq [d]$ is ϵ-sufficient with respect to a distribution $\mathcal{V}$ for f at $\mathbf{x}$ if*

$$\Delta_{\mathcal{V}}^{\textsf{suf}}(S, f, \mathbf{x}) \triangleq \rho(f(\mathbf{x}), f_S(\mathbf{x})) \leq \epsilon. \tag{1}$$

Furthermore, S is ϵ-super sufficient if all supersets $\widetilde{S} \supseteq S$ are ϵ-sufficient.

This notion of sufficiency is straightforward and aligns with P1. A subset S is ϵ-sufficient with respect to reference distribution $\mathcal{V}$ if, with $\mathbf{x}_S$ fixed, the average restricted prediction $f_S(\mathbf{x})$ is within ϵ from the original $f(\mathbf{x})$. Furthermore, S is ϵ-super sufficient if $\rho(f(\mathbf{x}), f_S(\mathbf{x})) \leq \epsilon$ and, $\forall \widetilde{S} \supseteq S$,

$\rho(f(\mathbf{x}), f_{\widetilde{S}}(\mathbf{x})) \leq \epsilon$. Namely, including more features in S keeps $f_S(\mathbf{x})$ ϵ close to $f(\mathbf{x})$. Note this definition aligns with P3, since the set $S = [d]$ is 0-sufficient (maximally sufficient). To find a small sufficient subset S of small cardinality $\tau > 0$, we can solve the following optimization problem:

$$\arg\min_{S \subseteq [d]} \ \Delta_{\mathcal{V}}^{\mathsf{suf}}(S, f, \mathbf{x}) \ \text{ subject to } \ |S| \leq \tau \tag{$\mathsf{P}_{\mathsf{suf}}$}$$

We will refer to this problem as the *sufficiency problem*, or ($\mathsf{P}_{\mathsf{suf}}$). Using analogous ideas, we also define necessity and formulate an optimization problem to find small necessary subsets.

Definition 2.2 (Necessity). *Let $\epsilon \geq 0$ and denote $\rho : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ to be metric on $\mathbb{R}$. A subset $S \subseteq [d]$ is ϵ-necessary with respect to a distribution $\mathcal{V}$ for f at $\mathbf{x}$ if*

$$\Delta_{\mathcal{V}}^{\mathit{nec}}(S, f, \mathbf{x}) \triangleq \rho(f_{S^c}(\mathbf{x}), f_{\emptyset}(\mathbf{x})) \leq \epsilon. \tag{2}$$

Furthermore, S is ϵ-super necessary if all supersets $\widetilde{S} \supseteq S$ are ϵ-necessary.

Here, a subset S is ϵ-necessary if marginalizing out the features in S with respect to $\mathcal{V}_S$, results in an average restricted prediction $f_{S^c}(\mathbf{x})$ that is ϵ close to $f_{\emptyset}(\mathbf{x})$ – the average baseline prediction of f over $\mathcal{V}_{[d]}$. Furthermore, S is ϵ-super necessary if $\rho(f_S(\mathbf{x}), f(\mathbf{x})) \leq \epsilon$ and, $\forall \widetilde{S} \supseteq S$, ϵ-necessary. Note, our definition of necessity differs from alternatives (Dhurandhar et al., 2018; Pawelczyk et al., 2020) which state that S is necessary if $\rho(f(\mathbf{x}), f_{S^c}(\mathbf{x})) \geq \Delta$ for some $\Delta > 0$. Our notion is more general in that it implies this condition. Intuitively, if $f_{\emptyset}(\mathbf{x})$ and $f(\mathbf{x})$ differ, and $f_{S^c}(\mathbf{x})$ is close to $f_{\emptyset}(\mathbf{x})$, then $f_{S^c}(\mathbf{x})$ and $f(\mathbf{x})$ will also differ. Furthermore, for $S = [d]$, we have $\Delta^{\mathsf{nec}}\mathcal{V}(S, f, \mathbf{x}) \triangleq \rho(f\emptyset(\mathbf{x}), f_{\emptyset}(\mathbf{x})) = 0$, indicating that $S = [d]$ is 0-necessary (maximally necessary) as desired. A detailed comparison of our approach with classical definitions, along with its advantages, is provided in the Appendix. To identify a ϵ-necessary subset S of small cardinality $\tau > 0$, one can solve the following optimization problem, which we refer to as the *necessity* problem or ($\mathsf{P}_{\mathsf{nec}}$).

$$\arg\min_{S \subseteq [d]} \ \Delta_{\mathcal{V}}^{\mathsf{nec}}(S, f, \mathbf{x}) \ \text{ subject to } \ |S| \leq \tau \tag{$\mathsf{P}_{\mathsf{nec}}$}$$

Having presented our definitions, we now discuss related works before presenting our main results.

3 Related Work

Notions of sufficiency, necessity, their duality and connections with other feature attribution methods have been studied to varying degrees. We comment on the main related works in this section.

Sufficiency. The notion of sufficient features has gained significant attention in recent research. Shih et al. (2018) explore a symbolic approach to explain Bayesian network classifiers and introduce prime implicant explanations, which are minimal subsets S that make features in the complement irrelevant to the prediction $f(\mathbf{x})$. For models represented by a finite set of first-order logic (FOL) sentences, Ignatiev et al. (2020) refer to prime implicants as abductive explanations (AXp's). For classifiers defined by propositional formulas and inputs with discrete features, Darwiche & Hirth (2020) refer to prime implicants as sufficient reasons and define a complete reason to be the disjunction of all sufficient reasons. They present efficient algorithms, leveraging Boolean circuits, to compute sufficient and complete reasons and demonstrate their use in identifying classifier dependence on protected features that should not inform decisions. For more complex models, Ribeiro et al. (2018) propose high-precision probabilistic explanations called anchors, which represent local, sufficient conditions. For $\mathbf{x}$ positively classified by f, Wang et al. (2021) propose a greedy approach to solve ($\mathsf{P}_{\mathsf{suf}}$), I Amoukou & Brunel (2022) extend this work to regression settings using tree-based models, and Fong & Vedaldi (2017) introduce the preservation method which relaxes S to $[0, 1]^d$.

Necessity. There has also been significant focus on identifying necessary features – those that, when altered, lead to a change in the prediction $f(\mathbf{x})$. For models expressible by FOL sentences, Ignatiev et al. (2019) define prime implicates as the minimal subsets that when changed, modify the prediction $f(\mathbf{x})$ and relate these to adversarial examples. For Boolean models predicting on samples $\mathbf{x}$ with discrete features, Ignatiev et al. (2020) and (Darwiche & Hirth, 2020) refer to prime implicates as contrastive explanations (CXp's) and necessary reasons, respectively. Beyond boolean functions, for $\mathbf{x}$ positively classified by a classifier f, Fong et al. (2019) relax S to $[0, 1]^d$ and propose the deletion method to approximately solve ($\mathsf{P}_{\mathsf{nec}}$).

Duality Between Sufficiency and Necessity. Dabkowski & Gal (2017) characterize the preservation and deletion methods as discovering the *smallest sufficient* and *destroying region* (SSR and SDR).

They propose combining the two but do not explore how solutions to this approach may differ from individual SSR and SDR solutions. Ignatiev et al. (2020) show that AXp's and CXp's are minimal hitting sets of another by using a hitting set duality result between minimal unsatisfiable and correction subsets. The result enables the identification of AXp's from CXp's and vice versa.

Sufficiency, Necessity, and General Feature Attribution Methods. Precise connections between sufficiency, necessity, and other popular feature attribution methods (such as Shapley values (Shapley, 1951; Chen et al., 2018b; Lundberg & Lee, 2017)) remains unclear. To our knowledge, Covert et al. (2021) provide the only work examining these approaches (Fong & Vedaldi, 2017; Fong et al., 2019; Dabkowski & Gal, 2017) in the context of general removal-based methods, i.e., methods that remove certain input features to evaluate different notions of importance. The work of Watson et al. (2021) is also relevant to our work, as it formalizes a connection between notions of sufficiency and Shapley values. With the specific payoff function defined as $v(S) = \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{S^c})]$, they show how each summand in the Shapley value measures the sufficiency of feature i to a particular subset.

4 Unifying Sufficiency and Necessity

Given a model f and sample $\mathbf{x}$, we can identify a small set of important features S by solving either ($\mathsf{P}_{\mathsf{suf}}$) or ($\mathsf{P}_{\mathsf{nec}}$). While both methods are popular (Kolek et al., 2022; Fong & Vedaldi, 2017; Bhalla et al., 2023; Yoon et al., 2018). identifying small sufficient or necessary subsets may not provide a complete picture of how f uses $\mathbf{x}$ to make a prediction. To see why, consider the following scenario: for a fixed $\tau > 0$, let S^* be a ϵ-sufficient solution to ($\mathsf{P}_{\mathsf{suf}}$), so that $|S^*| \leq \tau$ and $\Delta_{\mathcal{V}}^{\mathsf{suf}}(S, f, \mathbf{x}) \leq \epsilon$. While S^* is ϵ-sufficient, it can also be true that $\Delta_{\mathcal{V}}^{\mathsf{nec}}(S, f, \mathbf{x}) > \epsilon$ indicating S^* is **not** ϵ-necessary: indeed, this can simply happen when its complement, S^{c*}, contains important features. This scenario raises two questions: 1) How different are sufficient and necessary features? 2) How does varying the levels of sufficiency and necessity affect the optimal set of important features?

To answer these important questions (and avoid the scenario above) we propose studying a unification of ($\mathsf{P}_{\mathsf{suf}}$) and ($\mathsf{P}_{\mathsf{nec}}$).Consider $\Delta_{\mathcal{V}}^{\mathsf{uni}}(S, f, \mathbf{x}, \alpha) = \alpha \cdot \Delta_{\mathcal{V}}^{\mathsf{suf}}(S, f, \mathbf{x}) + (1-\alpha) \cdot \Delta_{\mathcal{V}}^{\mathsf{nec}}(S, f, \mathbf{x})$, a convex combination of $\Delta_{\mathcal{V}}^{\mathsf{suf}}(S, f, \mathbf{x})$ and $\Delta_{\mathcal{V}}^{\mathsf{nec}}(S, f, \mathbf{x})$, where $\alpha \in [0, 1]$ controls the extent to which S is sufficient vs. necessary. Our *unified problem*, ($\mathsf{P}_{\mathsf{uni}}$), can be expressed as:

$$\underset{S \subseteq [d]}{\arg\min} \ \Delta_{\mathcal{V}}^{\mathsf{uni}}(S, f, \mathbf{x}, \alpha) \ \text{ subject to } \ |S| \leq \tau \qquad (\mathsf{P}_{\mathsf{uni}})$$

When α is 1 or 0, $\Delta_{\mathcal{V}}^{\mathsf{uni}}(S, f, \mathbf{x}, \alpha)$ reduces to $\Delta_{\mathcal{V}}^{\mathsf{suf}}(S, f, \mathbf{x})$ or $\Delta_{\mathcal{V}}^{\mathsf{nec}}(S, f, \mathbf{x})$, respectively. In these extreme cases, S is only sufficient or necessary. In the remainder of this work we will theoretically analyze ($\mathsf{P}_{\mathsf{uni}}$), characterize its solutions, and provide different interpretations of what properties the solutions have through the lens of conditional independence and game theory. In the experimental section, we will show that solutions to ($\mathsf{P}_{\mathsf{uni}}$) provide insights that neither ($\mathsf{P}_{\mathsf{suf}}$) nor ($\mathsf{P}_{\mathsf{nec}}$) offer.

4.1 Solutions to the Unified Problem

We begin with a simple lemma that demonstrates why ($\mathsf{P}_{\mathsf{uni}}$) enforces both sufficiency and necessity.

Lemma 4.1. *Let $\alpha \in (0, 1)$. For $\tau > 0$, denote S^* to be a solution to* ($\mathsf{P}_{\mathsf{uni}}$) *for which $\Delta_{\mathcal{V}}^{uni}(S, f, \mathbf{x}, \alpha) = \epsilon$. Then, S^* is $\frac{\epsilon}{\alpha}$-sufficient and $\frac{\epsilon}{1-\alpha}$-necessary. Formally,*

$$0 \leq \Delta_{\mathcal{V}}^{suf}(S^*, f, \mathbf{x}) \leq \frac{\epsilon}{\alpha} \quad \textit{and} \quad 0 \leq \Delta_{\mathcal{V}}^{nec}(S^*, f, \mathbf{x}) \leq \frac{\epsilon}{1-\alpha}. \tag{3}$$

The proof of this result, and all others, is included Appendix A.1. This result illustrates that solutions to ($\mathsf{P}_{\mathsf{uni}}$) satisfy varying definitions of sufficiency and necessity. Furthermore, as α increases from 0 to 1, the solution shifts from being highly necessary to highly sufficient. In the following results, we will show *when* and *how* solutions to ($\mathsf{P}_{\mathsf{uni}}$) are similar (and different) to those of ($\mathsf{P}_{\mathsf{suf}}$) and ($\mathsf{P}_{\mathsf{nec}}$). To start, we present the following lemma, which will be useful in subsequent results.

Lemma 4.2. *For $0 \leq \epsilon < \frac{\rho(f(\mathbf{x}), f_{\emptyset}(\mathbf{x}))}{2}$, denote S^*_{suf} and S^*_{nec} to be ϵ-sufficient and ϵ-necessary sets. Then, if S^*_{suf} is ϵ-super sufficient or S^*_{nec} is ϵ-super necessary, we have $S^*_{suf} \cap S^*_{nec} \neq \emptyset$.*

This lemma demonstrates that, given ϵ-sufficient and necessary sets S^*_{suf} and S^*_{nec}, if either additionally satisfies the stronger notions of super sufficiency or necessity, they must share some features. This proves useful in characterizing a solution to ($\mathsf{P}_{\mathsf{uni}}$), which we now do in the following theorem.

Theorem 4.1. *Let $\tau_1, \tau_2 > 0$ and $0 \leq \epsilon < \frac{1}{2} \cdot \rho(f(\mathbf{x}), f_\emptyset(\mathbf{x}))$. Denote $S^*_{\textsf{suf}}$ and $S^*_{\textsf{nec}}$ to be ϵ-super sufficient and ϵ-super necessary solutions to* ($\mathsf{P}_{\textsf{suf}}$) *and* ($\mathsf{P}_{\textsf{nec}}$)*, respectively, such that $|S^*_{\textsf{suf}}| = \tau_1$ and $|S^*_{\textsf{nec}}| = \tau_2$. Then, there exists a set S^* such that*

$$\Delta^{\textsf{uni}}_{\mathcal{V}}(S^*, f, \mathbf{x}, \alpha) \leq \epsilon \quad \text{and} \quad \max(\tau_1, \tau_2) \leq |S^*| < \tau_1 + \tau_2. \tag{4}$$

Furthermore, if $S^*_{\textsf{suf}} \subseteq S^*_{\textsf{nec}}$ or $S^*_{\textsf{nec}} \subseteq S^*_{\textsf{suf}}$ then $S^* = S^*_{\textsf{nec}}$ or $S^* = S^*_{\textsf{suf}}$ respectively.

This result demonstrates that when there are ϵ-super sufficient and ϵ-super necessary solutions to ($\mathsf{P}_{\textsf{suf}}$) and ($\mathsf{P}_{\textsf{nec}}$), then one can identify a set S^* with small $\Delta^{\textsf{uni}}$. As an example, consider features that are ϵ-super sufficient, $S^*_{\textsf{suf}}$. If we have domain knowledge that $S^*_{\textsf{suf}} \subseteq S^*_{\textsf{nec}}$, and $S^*_{\textsf{nec}}$ is ϵ-super necessary, then $S^*_{\textsf{nec}}$ will have a small $\Delta^{\textsf{uni}}$ Conversely, if we know that $S^*_{\textsf{suf}}$ is ϵ-super necessary along with being a subset of ϵ-super sufficient set $S^*_{\textsf{suf}}$, then $S^*_{\textsf{suf}}$ will have a small $\Delta^{\textsf{uni}}$.

5 Two Perspectives of the Unified Approach

In the previous section, we characterized solutions to ($\mathsf{P}_{\textsf{uni}}$) and their connections to those of ($\mathsf{P}_{\textsf{suf}}$) and ($\mathsf{P}_{\textsf{nec}}$). To further motivate and the unified approach, we now offer two alternative perspectives of our framework through the lens of conditional independence and Shapley values.

5.1 A Conditional Independence Perspective

Here we demonstrate how sufficiency, necessity, and their unification, can be understood as conditional independence relations between features $\mathbf{X}$ and label Y.

Corollary 5.1. *Suppose $\forall S \subseteq [d]$, $\mathcal{V}_S = p(\mathbf{X}_S | \mathbf{X}_{S^c} = \mathbf{x}_{S^c})$. Let $\alpha \in (0, 1)$, $\epsilon \geq 0$, and denote $\rho : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ to be a metric. Furthermore, for $\tau > 0$ and $f(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$, let S^* be a solution to* ($\mathsf{P}_{\textsf{uni}}$) *such that $\Delta^{\textsf{uni}}_{\mathcal{V}}(S, f, \mathbf{x}, \alpha) = \epsilon$. Then, S^* satisfies the follow conditional independencies,*

$$\rho\left(\mathbb{E}[Y \mid \mathbf{x}],\, \mathbb{E}[Y \mid \mathbf{X}_{S^*} = \mathbf{x}_{S^*}]\right) \leq \frac{\epsilon}{\alpha} \quad \text{and} \quad \rho\left(\mathbb{E}[Y \mid \mathbf{X}_{S^*_c} = \mathbf{x}_{S^*_c}],\, \mathbb{E}[Y]\right) \leq \frac{\epsilon}{1-\alpha}. \tag{5}$$

The assumption in this corollary is that, $\forall\, S \subseteq [d]$, $f_S(\mathbf{x})$ is evaluated using the conditional distribution $p(\mathbf{X}_{S^c} \mid \mathbf{X}_S = \mathbf{x}_S)$ as the reference distribution $\mathcal{V}_S$. Given the recent advancements in generative models (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021), this assumption is (approximately) reasonable in many practical settings, as we will demonstrate in our experiments. For this particular $\mathcal{V}_S$, the result shows that minimizing ($\mathsf{P}_{\textsf{uni}}$) with model $f(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$ identifies an S^* that approximately satisfies two conditional independence properties. First, S^* is sufficient as conditioning on S^* leaves the complement S^{c*} with minimal additional information about Y. Second, S^* is necessary because when we solely rely on the complement S^{c*}, the information gained about Y is minimal and similar to $\mathbb{E}[Y = 1]$.

5.2 A Shapley Value Perspective

In the previous section, we detailed the conditional independence relations being optimized for when solving ($\mathsf{P}_{\textsf{uni}}$). We now present an arguably less intuitive result that shows that solving ($\mathsf{P}_{\textsf{uni}}$) is equivalent to maximizing the lower bound of the Shapley value. Before presenting our result, we provide a brief background on this game-theoretic quantity.

Shapley Values. Shapley values use game theory to measure the importance of players in a game. Let the tuple $([n], v)$ represent a cooperative game with players $[n] = \{1, 2, \ldots, n\}$ and denote a characteristic function $v(S) : \mathcal{P}([n]) \to \mathbb{R}$, Then, the Shapley value (Shapley, 1951) for player j in the game $([n], v)$ is $\phi^{\textsf{shap}}_j([n], v) = \sum_{S \subseteq [n] \setminus \{j\}} w_S \cdot [v(S \cup \{j\}) - v(S)]$ where $w_S = \frac{|S|!(n-|S|-1)!}{n!}$. In the context of XAI, Shapley values are widely used to measure local feature importance by treating input features as players in a game (Covert et al., 2020; Teneggi et al., 2022; Chen et al., 2018b; Lundberg & Lee, 2017). Given a sample $\mathbf{x}$ and a model f, the importance of x_j to the prediction $f(\mathbf{x})$ is measured by computing $\phi^{\textsf{shap}}_j$ for a game $([d], v)$, where $v(S)$ quantifies how the features in S contribute to $f(\mathbf{x})$. Different choices of $v(S)$ can be found in (Lundberg & Lee, 2017; Sundararajan & Najmi, 2020; Watson et al., 2024). Although computing $\phi^{\textsf{shap}}_j$ is computationally intractable, several practical methods for estimation have been developed (Chen et al., 2023; Teneggi et al., 2022; Zhang et al., 2023; Lundberg et al., 2020). While Shapley values are popular across various domains (Moncada-Torres et al., 2021; Zoabi et al., 2021; Liu et al., 2021), few works, aside from Watson et al. (2021), explore their connections to sufficiency and necessity.

With this background, we now present our result. Recall solving (P_{uni}) finds a small subset S with low $\Delta_{\mathcal{V}}^{\text{uni}}(S, f, \mathbf{x}, \alpha)$. Notice that ($P_{\text{uni}}$) naturally *partitions* the features into two sets, S and S^c. In the following theorem we demonstrate that finding a small S with minimal $\Delta_{\mathcal{V}}^{\text{uni}}(S, f, \mathbf{x}, \alpha)$ is equivalent to maximizing a lower bound on the Shapley value in a two player game.

Theorem 5.1. *Consider an input $\mathbf{x}$ for which $f(\mathbf{x}) \neq f_\emptyset(\mathbf{x})$. Denote by $\Lambda_d = \{S, S^c\}$ the partition of $[d] = \{1, 2, \ldots, d\}$, and define the characteristic function to be $v(S) = -\rho(f(\mathbf{x}), f_S(\mathbf{x}))$. Then,*

$$\phi_S^{\text{shap}}(\Lambda_d, v) \geq \rho(f(\mathbf{x}), f_\emptyset(\mathbf{x})) - \Delta_{\mathcal{V}}^{\text{uni}}(S, f, \mathbf{x}, \alpha). \tag{6}$$

This result motivates minimizing $\Delta_{\mathcal{V}}^{\text{uni}}(S, f, \mathbf{x}, \alpha)$ via a game-theoretic interpretation. The tuple (Λ_d, v) specifies a game, and since there are 2^{d-1} ways to partition $[d]$ into 2 subsets, there are 2^{d-1} games. The inequality above holds for each of them. Thus, Theorem 5.1 implies that finding the S with minimal $\Delta_{\mathcal{V}}^{\text{uni}}(S, f, \mathbf{x}, \alpha)$ is equivalent to identifying the game (i.e. partition) (Λ_d, v) in which S has the largest lower bound on its Shapley value.

6 Solving the Unified Problem

Before presenting our results, we briefly discuss different approaches to solving (P_{uni}). In general, this problem is NP-hard however, in certain settings, one can efficiently compute exact solutions or use tractable relaxations, (Kolek et al., 2022; Fong et al., 2019; Linder et al., 2022) to approximate solutions. We present these general approaches here, and defer details to Appendix A.2.

Exhaustive Search. When the feature space dimension, d, or choice of $\tau \in \mathbb{Z}_{>0}$ is small an exhaustive search can compute exact solutions to (P_{uni}) by evaluating $\Delta_{\mathcal{V}}^{\text{uni}}(S, f, \mathbf{x}, \alpha)$ for all $\binom{d}{\tau}$ subsets S of cardinality τ and selecting the minimizer.

Instance-wise Optimization. When d is large, rendering (P_{uni}) intractable, one can generate approximate solutions by solving the relaxed problem[1]

$$\underset{S \subseteq [0,1]^d}{\arg\min} \;\; \Delta_{\mathcal{V}}^{\text{uni}}(S, f, \mathbf{x}, \alpha) + \lambda_1 \cdot ||S||_1 + \lambda_{\text{TV}} \cdot ||S||_{TV}. \tag{7}$$

This type of approach is often used in computer vision and natural language problems (Fong et al., 2019; Kolek et al., 2022; Linder et al., 2022) to generate instance-specific solutions.

Parametric Model Approach. Another we approach we take to generate solutions to (P_{uni}) is to learn models $g_\theta : \mathcal{X} \mapsto [0, 1]^d$ that (approximately) solve the following optimization problem:

$$\underset{\theta \in \Theta}{\arg\min} \;\; \underset{\mathbf{X} \sim \mathcal{D}_{\mathcal{X}}}{\mathbb{E}} \left[\Delta_{\mathcal{V}}^{\text{uni}}(g_\theta(\mathbf{X}), f, \mathbf{X}, \alpha) + \lambda_1 \cdot ||g_\theta(\mathbf{X})||_1 + \lambda_{\text{TV}} \cdot ||g_\theta(\mathbf{X})||_{\text{TV}} \right]. \tag{8}$$

With these models, an approximate solution can be computed via $g_\theta(\mathbf{x})$. This method is popular (Chen et al., 2018a; Yoon et al., 2018; Linder et al., 2022), as it handles highly structured data well and requires training only one model, rather than repeatedly solving Eq. (7) for each sample.

7 Experiments

We demonstrate our theoretical findings in multiple settings of increasingly complexity: two tabular data tasks (on synthetic data and the US adult income dataset (Ding et al., 2021)) and two high-dimensional image classification tasks using the RSNA 2019 Brain CT Hemorrhage Challenge (Flanders et al., 2020) and CelebA-HQ datasets (Lee et al., 2020)

7.1 Tabular Data

With the following tabular data settings, we demonstrate how the specific trade-off between sufficiently and necessity can greatly alter the solutions to (P_{uni}). To do so, we compute exact solutions via exhaustive search to (P_{uni}) for varying levels of sufficiency vs. necessity and multiple size constraints. We learn a predictor f and, for 100 new samples, solve (P_{uni}) for $\tau \in \{3, 6, 9\}$ and $\alpha \in [0, 1]$, with $\rho(a, b) = |a - b|$ and $\mathcal{V}_S = p(\mathbf{X}_S \mid \mathbf{X}_{S^c} = \mathbf{x}_{S^c})$. For a fixed τ and sample $\mathbf{x}$, we denote $S^*_{\alpha_i}$ to be a solution to (P_{uni}) for α_i. It is represented as a binary vector $s \in \{0, 1\}^{10}$, where $s_j = 1$ if $j \in S^*_{\alpha_i}$ and 0 otherwise. To analyze the stability of $S^*_{\alpha_i}$ as sufficiency and necessity vary, we report the normalized average Hamming distance (Hamming, 1950) between $S^*_{\alpha_i}$ and S^*_0 (with 95% confidence intervals) as a function of α.

[1]Here, λ_1, $||S||_1$ and λ_{TV}, $||S||_{TV}$ are the ℓ_1 and Total Variation norms and hyperparamters, respectively, promoting sparsity and smoothness.

7.1.1 Linear Regression

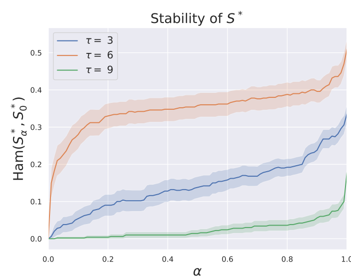
We begin with a regression example. Features are distributed as $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}\mathbf{A}^{\mathbf{T}})$ with $\boldsymbol{\mu} = [2^i]_{i=1}^d$ and $\mathbf{A}_{i,j} \sim U(0,1)$. The response is $Y = \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\epsilon}$, with $\boldsymbol{\beta} = 32 \cdot [2^{-i}]_{i=1}^d$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$. We fix $d = 10$ and use the model $f(\mathbf{X}) = \hat{\boldsymbol{\beta}}^T\mathbf{X}$, where $\hat{\boldsymbol{\beta}}$ is the least squares solution.

Stability of Unified Solutions. Fig. 1a shows that when solutions are constrained to be small ($\tau = 3$), increasing α to enforce greater sufficiency results in a steady increase inHamming distance, indicating that the solutions $S^*_{\alpha_i}$ are consistently changing. When larger solutions are allowed ($\tau = 6$), $S^*_{\alpha_i}$ rapidly changes with the introduction of sufficiency, as seen by the initial steep rise in Hamming distance. However, as α continues to increase, this distance grows more gradually. Lastly, when the solution size approaches the dimension of the feature space ($\tau = 9$), small to medium levels of sufficiency do not significantly alter $S^*_{\alpha_i}$. However, high levels of sufficiency ($\alpha > 0.8$) lead to extreme changes in the solutions, as shown by a sharp increase in Hamming distance.

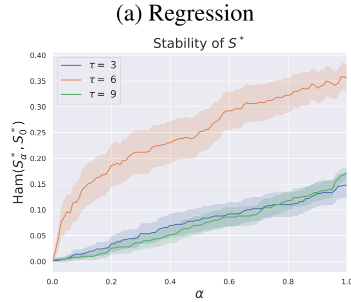
7.1.2 American Community Survey Income (ACSIncome)

We use the ACSIncome dataset for California, including 10 demographic and socioeconomic features such as age, education, occupation, and geographic region. We train a Random Forest classifier to predict whether an individual's annual income exceeds \$50K, achieving a test accuracy $\approx 81\%$.

Stability of Unified Solutions. Fig. 1b shows that when solutions are forced to be small ($\tau = 3$), increasing α to enforce sufficiency results in a steady increase in Hamming distance, indicating the solutions $S^*_{\alpha_i}$ are changing. For larger solutions ($\tau = 6$), $S^*_{\alpha_i}$ changes significantly when low levels sufficiency are required, indicated by initial rise in the Hamming distance. As α continues to increase, the Hamming distance grows more gradually. Interestingly, when the size is close to feature space's dimensionality ($\tau = 9$), the Hamming distance exhibits a behavior similar to that observed for $\tau = 3$. In conclusion, both examples show that the optimal feature set can vary depending on the size constraint and balance between sufficiency and necessity.



(a) Regression



(b) ACSIncome

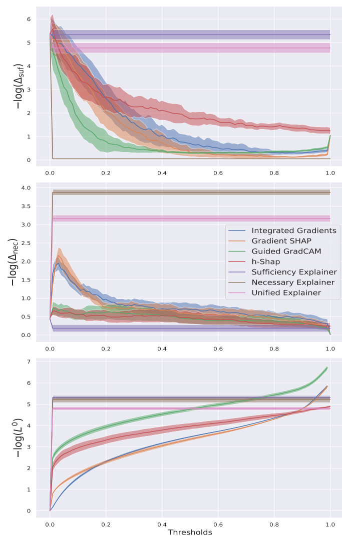
Figure 1: Stability of ($\mathsf{P}_{\mathsf{uni}}$) Solutions

7.2 Image Classification

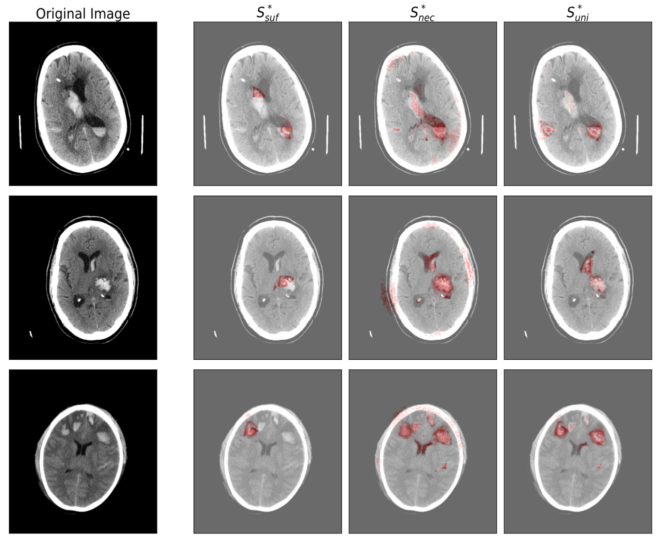
The following two experiments explore high dimensional image classification tasks. The features are pixel values and so a subset S corresponds to a binary mask identifying important pixels. Since solving ($\mathsf{P}_{\mathsf{suf}}$), ($\mathsf{P}_{\mathsf{nec}}$), or ($\mathsf{P}_{\mathsf{uni}}$) is intractable here, we use two methods, the per-sample and model based approach in Eqs. (7) and (8) to identify sufficient and necessary masks. These experiments serve two purposes. First, they will analyze the ability popular explanation methods–including Integrated Gradients (Sundararajan et al., 2017), GradientSHAP (Lundberg & Lee, 2017), Guided GradCAM (Selvaraju et al., 2017), and h-Shap (Teneggi et al., 2022)–to identify small sufficient and necessary subsets. To ensure consistent analysis, all attribution scores are normalized to the interval $[0, 1]$. This is done by setting the top 1% of nonzero scores to 1 and dividing the remaining by the minimum score from the top 1% nonzero scores, which is common practice (Kokhlikyan et al., 2020). Binary masks are then generated by thresholding the normalized scores using thresholds $t \in (0, 1)$. For a test set of images and normalized attribution scores, we report the average (across all binary masks) $-\log(\Delta^{\mathsf{suf}})$, $-\log(\Delta^{\mathsf{nec}})$, and $-\log(L^0)$ where L^0 is the relative size of S for $t \in (0, 1)$ to analyze the sufficiency, necessity and size of the explanations. The second objective of these experiments is to understand and visualize the similarities and differences between sufficient and necessary sets.

7.2.1 RSNA CT Hemorrhage

We use the RSNA 2019 Brain CT Hemorrhage Challenge dataset comprised of 752,803 scans. Each scan is annotated by expert neuroradiologists with the presence and type(s) of hemorrhage (i.e., epidural, intraparenchymal, intraventricular, subarachnoid, or subdural). We use a ResNet18 (He et al., 2016) classifier that was pretrained on this data (Teneggi et al., 2022). Since the dataset



(a) Comparison of different methods.



(b) S^*_{suf}, S^*_{nec} and S^*_{uni} for various CT scans.

Figure 2: Experimental results on the RSNA dataset.

consists of highly complex and diverse images, we employ the per-example approach in Eq. (7) with $\alpha \in \{0, 0.5, 1\}$ to learn sufficient and necessary masks. Further details are in Appendix A.2.

Comparison of Post-hoc Interpretability Methods. For a set of 20 images positively classified by the ResNet model, we apply multiple post-hoc interpretability methods, as well as compute sufficient and necessary masks by solving (7). The results in Fig. 2a show that for thresholds $t < 0.1$, many methods identify sufficient sets smaller in size than the sufficient and unified explainer, as indicated by their large values of $-\log(\Delta^{\text{suf}})$ and smaller values of $-\log(L^0)$. However, for $t > 0.1$, only the sufficient and unified explainer identify sufficient sets of a constant small size. Importantly, *no methods, besides the necessity and unified explainers, identify necessary sets*. Furthermore, as expected, the sufficient explainer does not identify necessary sets and vice versa. The unified explainer, as expected, identifies a sufficient and necessary set (at the cost of a larger set). In conclusion, while off-the-shelf methods can identify sufficient, they do not identify necessary sets for small thresholds.

Sufficiency vs. Necessity. In Fig. 2b we visualize the sufficient and necessary features in various CT scans. The first observation is that sufficient subsets do not provide a complete picture of which features are important. Notice for all the CT scans, a sufficient set, S^*_{suf} highlights one or two, but never all, brain hemorrhages in the scans. For example, in the last row, S^*_{suf} only contains the right frontal lobe parenchymal hemorrhages, which happens to be one of the larger hemorrhages present. On the other hand, necessary sets, S^*_{nec}, contain parts of, sometimes entirely, *all* hemorrhages in the scans. In the last row, S^*_{nec} contains all multifocal parenchymal hemorrhages in both right and left frontal lobes, because when all these regions are masked, the model yields a prediction ≈ 0.64– the prediction of the model on the mean image. Finally, notice in the 2nd and 3rd columns that S^*_{nec} and S^*_{uni} are nearly identical, which precisely demonstrate Lemma 4.1 and Theorem 4.1 in practice. First, since S^*_{suf} is super sufficient, S^*_{suf} and S^*_{nec}, share common features. Second, visually $S^*_{\text{suf}} \subseteq S^*_{\text{nec}}$ holds approximately and so $S^*_{\text{nec}} = S^*_{\text{uni}}$. Through this experiment we are able to highlight the differences between sufficient and necessary sets, show how each contain important and complementary information, and demonstrate our theory holding in real world settings.

7.2.2 CelebA-HQ

We use a modified version of the CelebA-HQ dataset (Karras, 2017) that contains 30,000 celebrity faces resized to 256×256 pixels. We train a ResNet18 to classify whether a celebrity is smiling, achieving a test accuracy $\approx 94\%$ and use the model based approach via solving Eq. (8) to generate sufficient and necessary masks. Given the structured nature of the dataset and the similarity of features across images, we use the model approach because it prevents overfitting to spurious signals

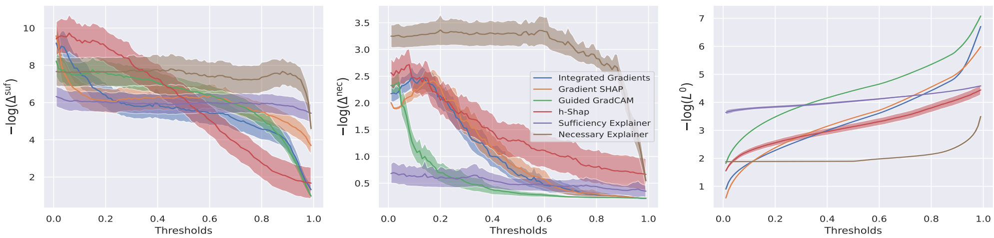


Figure 3: Comparison of different methods on the CelebAHQ dataset.

(Linder et al., 2022), an issue that can arise with per-example methods. Implementation details and hyperparameter settings are included in Appendix A.2.

Comparison of Post-hoc Interpretability Methods. For a set of 100 images labeled with a smile and correctly classified by the ResNet classifier, we apply multiple post-hoc interpretability methods and our sufficient and necessary explainers to identify important features associated with smiling. The results in Fig. 3 illustrate that for a wide range of thresholds $t \in [0, 1]$, many methods identify sufficient subsets, as $-\log(\Delta^{\textsf{suf}})$ for many of them is comparable to that of the sufficient explainer. The necessary explainer, in fact, identifies subsets that are more sufficient than those found by the sufficient explainer. The reason is that the sufficient explainer identifies subsets that are, on average, smaller for all $t \in [0, 1]$, while the necessary explainer finds subsets that are constant in size for all $t \in [0, 1]$ but slightly larger since, to be necessary, they must contain more features that provide additional information about the label. For other methods, as t increases, subset size decreases, and the sufficiency and necessity of the solutions decline. Meanwhile, the necessary explainer naturally identifies necessary subsets, indicated by large $-\log(\Delta^{\textsf{nec}})$, whereas other methods fail to do so. In conclusion, many methods can identify sufficient sets, but not necessary ones and directly optimizing for these criterion leads to identifying small, constant-sized subsets across thresholds.

Sufficiency vs. Necessity. In Fig. 4, we see how sufficient subsets alone may overlook important features, while solutions to ($\text{P}_{\textsf{uni}}$) offer deeper insights. As stated earlier, the sufficient explainer identifies sets that are sufficient but not necessary. On the other hand, the necessary explainer has high $-\log(\Delta^{\textsf{suf}})$ and $-\log(\Delta^{\textsf{nec}})$, indicating that it identifies sufficient *and* necessary set, meaning they also serve as solutions to ($\text{P}_{\textsf{uni}}$). In Fig. 4, we visualize the reasons for this phenomena. Notice that $S^*_{\textsf{suf}}$ precisely highlights (only) the smile. When $S^*_{\textsf{suf}}$ is fixed, one can generate new images (as done in (Zhang et al., 2023)) for which the model produces the same predictions as it did for the original image (a smile). On the other hand, we also see why $S^*_{\textsf{suf}}$ is *not* necessary: we can fix the complement $(S^*_{\textsf{suf}})_c$ and, since there are important features in it, a smile is consistently generated, and the model produces the same prediction on these images as it did on the original. Conversely solutions to ($\text{P}_{\textsf{nec}}$) (also solutions to ($\text{P}_{\textsf{uni}}$) here) generate different explanations that provide a more complete picture of feature importance. Notice that $S^*_{\textsf{nec}}$ is sufficient because $S^*_{\textsf{suf}} \subseteq S^*_{\textsf{nec}}$, with the additional features mainly being the dimples and eyes, which aid in determining the presence of a smile. More importantly, Fig. 5 illustrates why $S^*_{\textsf{nec}}$ is necessary: when we fix the complement of $S^*_{\textsf{nec}}$ and generate new samples, half of the faces lack a smile, leading the model f to predict no smile. Additional images and details on sample generation are in Appendices A.2 and A.4.

8 Limitations & Broader Impacts

While this work provides a novel theoretical contribution to the XAI community, there are some limitations that require careful discussion. The choice of reference distribution $\mathcal{V}_S$ determines the characteristics of sufficient and necessary explanations. For instance, only with the true conditional data distribution can one obtain the conditional independence results that our theory provides. Naturally, there are computational trade-offs that must be carefully studied; the ability to learn and sample from accurate conditional distributions to generate explanations with clear statistical meaning comes with a computational and statistical cost, particularly in high-dimensional settings. Thus, a key direction for future work is to explore the impact of different reference distributions and provide a principled framework for selecting a $\mathcal{V}_S$ that balances practical utility and computational feasibility.

Another relevant question is how well our proposed notions align with human intuition. While we aim to understand which features are sufficient and necessary *for a given predicted model*, these explanations may not always correspond to how humans perceive importance (since model might

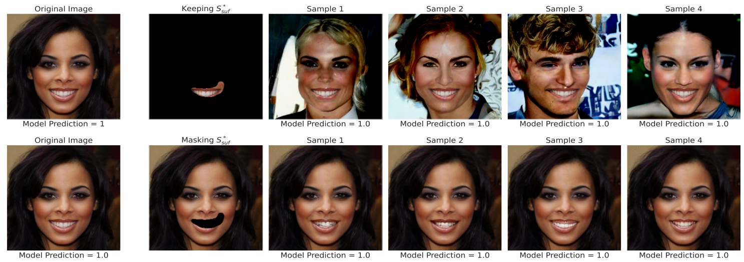


Figure 4: Images and model predictions by fixing and masking the sufficient subset $S^*_{\textsf{suf}}$

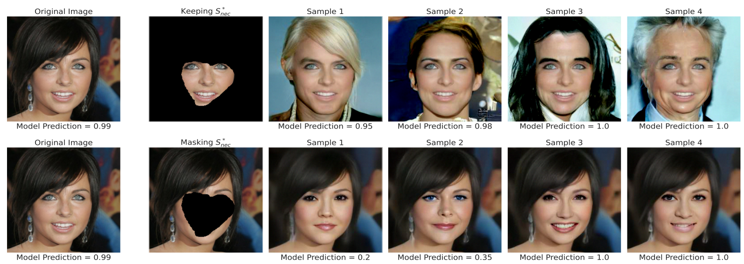


Figure 5: Images and model predictions by fixing and masking the necessary subset $S^*_{\textsf{nec}}$

use different features to solve a task). This can be an issue in settings where interpretability is essential for trust and accountability, such as in healthcare. On the one hand, our approach can provide useful insights to further evaluate models (e.g. by verifying if the sufficient and necessary features employed by models correlate with the correct ones as informed by human experts). On the other hand, bridging the gap between our mathematical definitions of sufficiency and necessity and other human notions of importance is an area for further investigation. User studies, along with collaboration with domain experts, will be critical in determining how our formal notions of sufficiency and necessity can be adapted or extended to better meet real-world interpretability needs.

Finally, the societal impact of this work warrants discussion. While we offer a rigorous framework to understand model predictions, these are oblivious to notions of demographic bias (Hardt et al., 2016; Feldman et al., 2015; Bharti et al., 2024). There is a risk that an "incorrect" choice of generating a sufficient vs. necessary explanation could reinforce biases or obscure the causal reasons behind predictions. Future work will study when and how our framework can incorporate these biases.

9 Conclusion

This work formalizes notions of sufficiency and necessity as tools to evaluate feature importance and explain model predictions. We demonstrate that sufficient and necessary explanations, while insightful, often provide incomplete while complementary answers to model behavior. To address this limitation, we propose a unified approach that offers a new and more nuanced understanding of model behavior. Our unified approach expands the scope of explanations and reveals trade-offs between sufficiency and necessity, giving rise to new interpretations of feature importance. Through our theoretical contributions, we present conditions under which sufficiency and necessity align or diverge, and provide two perspectives of our unified approach through the lens of conditional independence and Shapley values. Our experimental results support our theoretical findings, providing examples of how adjusting sufficiency-necessity trade-off via our unified approach can uncover alternative sets of important features that would be missed by focusing solely on sufficiency or necessity. Furthermore, we evaluate common post-hoc interpretability methods showing that many fail to reliably identify features that are necessary or sufficient. In summary, our work contributes to a more complete understanding of feature importance through sufficiency and necessity. We believe, and hope, our framework holds potential for advancing the rigorous interpretability of ML models.

References

Usha Bhalla, Suraj Srinivas, and Himabindu Lakkaraju. Verifiable feature attributions: A bridge between post hoc explainability and inherent interpretability. *Advances in neural information* *processing systems*, 2023.

Beepul Bharti, Paul Yi, and Jeremias Sulam. Estimating and controlling for equalized odds via sensitive attribute predictors. *Advances in neural information processing systems*, 36, 2024.

Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, pp. 1–12, 2023.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International conference on machine learning*, pp. 883–892. PMLR, 2018a.

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*, 2018b.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223, 2020.

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.

Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017.

Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *ECAI 2020*, pp. 712–720. IOS Press, 2020.

Adnan Darwiche and Chunxi Ji. On the computation of necessary and sufficient explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5582–5591, 2022.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31, 2018.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3):e190211, 2020.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2950–2958, 2019.

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.

Richard W Hamming. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Salim I Amoukou and Nicolas Brunel. Consistent sufficient explanations and minimal local rules for explaining the decision of any classifier or regressor. *Advances in Neural Information Processing Systems*, 35:8027–8040, 2022.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. *Advances in neural information processing systems*, 32, 2019.

Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva. From contrastive to abductive explanations and back again. In *International Conference of the Italian Association for Artificial Intelligence*, pp. 335–355. Springer, 2020.

Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *International Conference on Artificial Intelligence and Statistics*, pp. 1459–1467. PMLR, 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.

Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. URL `https://arxiv.org/abs/2009.07896`.

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.

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

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Johannes Linder, Alyssa La Fleur, Zibo Chen, Ajasja Ljubetič, David Baker, Sreeram Kannan, and Georg Seelig. Interpreting neural networks for biological sequences by learning stochastic masks. *Nature machine intelligence*, 4(1):41–54, 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.

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

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 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.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, pp. 809–818. PMLR, 2020.

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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.

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

Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. *arXiv preprint arXiv:1805.03364*, 2018.

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.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

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*, 45(4):4494–4503, 2022.

Jacopo Teneggi, Beepul Bharti, Yaniv Romano, and Jeremias Sulam. Shap-xrt: The shapley value meets conditional independence testing. *Transactions on Machine Learning Research*, 2023.

The White House. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, 2023.

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.

Eric Wang, Pasha Khosravi, and Guy Van den Broeck. Probabilistic sufficient explanations. *arXiv preprint arXiv:2105.10118*, 2021.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, and Shir. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972): 47–60, August 2023.

David Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. Explaining predictive uncertainty with information theoretic shapley values. *Advances in Neural Information Processing Systems*, 36, 2024.

David S. Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. Local explanations via necessity and sufficiency: unifying theory and practice. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1382–1392. PMLR, 27–30 Jul 2021.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Invase: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2018.

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

Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi Jaakkola, and Shiyu Chang. Towards coherent image inpainting using denoising diffusion implicit models. In *International Conference on Machine Learning*, pp. 41164–41193. PMLR, 2023.

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.

A APPENDIX

A.1 PROOFS

A.1.1 PROOF OF LEMMA 4.1

Lemma 4.1. Let $\alpha \in (0,1)$. For $\tau > 0$, denote S^* to be a solution to ($\mathrm{P}_{\mathsf{uni}}$) for which $\Delta_{\mathcal{V}}^{\mathsf{uni}}(S^*, f, \mathbf{x}, \alpha) = \epsilon$. Then, S^* is $\frac{\epsilon}{\alpha}$-sufficient and $\frac{\epsilon}{1-\alpha}$-necessary. Formally,

$$0 \leq \Delta_{\mathcal{V}}^{\mathsf{suf}}(S^*, f, \mathbf{x}) \leq \frac{\epsilon}{\alpha} \quad \text{and} \quad 0 \leq \Delta_{\mathcal{V}}^{\mathsf{nec}}(S^*, f, \mathbf{x}) \leq \frac{\epsilon}{1-\alpha}. \tag{9}$$

Proof. Let $\tau > 0$ and $\alpha \in (0,1)$ and denote S^* to be a solution to ($\mathrm{P}_{\mathsf{uni}}$) such that

$$\Delta_{\mathcal{V}}^{\mathsf{uni}}(S^*, f, \mathbf{x}, \alpha) = \epsilon. \tag{10}$$

Then, by definition of being a solution to ($\mathrm{P}_{\mathsf{uni}}$),

$$|S^*| \leq \tau. \tag{11}$$

Furthermore, recall that

$$\Delta_{\mathcal{V}}^{\mathsf{uni}}(S^*, f, \mathbf{x}, \alpha) = \alpha \cdot \Delta_{\mathcal{V}}^{\mathsf{suf}}(S^*, f, \mathbf{x}) + (1-\alpha) \cdot \Delta_{\mathcal{V}}^{\mathsf{nec}}(S^*, f, \mathbf{x}) \tag{12}$$

which implies

$$\begin{aligned}
\alpha \cdot \Delta_{\mathcal{V}}^{\mathsf{suf}}(S^*, f, \mathbf{x}) &= \epsilon - (1-\alpha) \cdot \Delta_{\mathcal{V}}^{\mathsf{nec}}(S^*, f, \mathbf{x}) && (13)\\
&\leq \epsilon \qquad\qquad ((1-\alpha),\ \Delta_{\mathcal{V}}^{\mathsf{nec}}(S^*, f, \mathbf{x}) \geq 0) && (14)\\
\implies \Delta_{\mathcal{V}}^{\mathsf{suf}}(S^*, f, \mathbf{x}) &\leq \frac{\epsilon}{\alpha}. && (15)
\end{aligned}$$

Similarly,

$$\begin{aligned}
(1-\alpha) \cdot \Delta_{\mathcal{V}}^{\mathsf{nec}}(S^*, f, \mathbf{x}) &= \epsilon - \alpha \cdot \Delta_{\mathcal{V}}^{\mathsf{suf}}(S^*, f, \mathbf{x}) && (16)\\
&\leq \epsilon \qquad\qquad (\alpha,\ \Delta_{\mathcal{V}}^{\mathsf{suf}}(S^*, f, \mathbf{x}) \geq 0) && (17)\\
\implies \Delta_{\mathcal{V}}^{\mathsf{nec}}(S^*, f, \mathbf{x}) &\leq \frac{\epsilon}{1-\alpha}. && (18)
\end{aligned}$$

□

A.1.2 PROOF OF LEMMA 4.2

Lemma 4.2. For $0 \leq \epsilon < \frac{\rho(f(\mathbf{x}), f_{\emptyset}(\mathbf{x}))}{2}$, denote S^*_{suf} and S^*_{nec} to be ϵ-sufficient and ϵ-necessary sets. Then, if S^*_{suf} is ϵ-super sufficient or S^*_{nec} is ϵ-super necessary,

$$S^*_{\mathsf{suf}} \cap S^*_{\mathsf{nec}} \neq \emptyset. \tag{19}$$

Proof. We will prove the result via contradiction. First recall that,

$$f_S(\mathbf{x}) = \mathop{\mathbb{E}}_{\mathbf{X}_{S^c} \sim \mathcal{V}_{S^c}} [f(\mathbf{x}_S, \mathbf{X}_{S^c})] \tag{20}$$

and, for any metric $\rho : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$,

$$\Delta_{\mathcal{V}}^{\mathsf{suf}}(S, f, \mathbf{x}) \triangleq \rho(f(\mathbf{x}), f_S(\mathbf{x})) \tag{21}$$

$$\Delta_{\mathcal{V}}^{\mathsf{nec}}(S, f, \mathbf{x}) \triangleq \rho(f_{S^c}(\mathbf{x}), f_{\emptyset}(\mathbf{x})). \tag{22}$$

Since ρ is a metric on $\mathbb{R}$, it satisfies the triangle inequality. Thus, for $a, b, c \in \mathbb{R}$

$$\rho(a, c) \leq \rho(a, b) + \rho(b, c). \tag{23}$$

Now, let S^*_{suf} be ϵ-super sufficient and suppose

$$S^*_{\mathsf{suf}} \cap S^*_{\mathsf{nec}} = \emptyset. \tag{24}$$

This implies

$$S^*_{\textsf{suf}} \subseteq (S^*_{\textsf{nec}})_c. \tag{25}$$

Subsequently, since $S^*_{\textsf{suf}}$ is ϵ-super sufficient,

$$\Delta^{\textsf{suf}}_{\mathcal{V}}((S^*_{\textsf{nec}})_c, f, \mathbf{x}) \leq \epsilon. \tag{26}$$

As a result, observe

$$\begin{aligned}
\rho(f(\mathbf{x}), f_\emptyset(\mathbf{x})) &\leq \rho(f(\mathbf{x}), f_{(S^*_{\textsf{nec}})_c}(\mathbf{x})) + \rho(f_{(S^*_{\textsf{nec}})_c}(\mathbf{x}), f_\emptyset(\mathbf{x})) && \text{triangle inequality} && (27)\\
&= \Delta^{\textsf{suf}}_{\mathcal{V}}((S^*_{\textsf{nec}})_c, f, \mathbf{x}) + \Delta^{\textsf{nec}}_{\mathcal{V}}((S^*_{\textsf{nec}})_c, f, \mathbf{x}) && && (28)\\
&\leq \epsilon + \Delta^{\textsf{nec}}_{\mathcal{V}}((S^*_{\textsf{nec}})_c, f, \mathbf{x}) && S^*_{\textsf{suf}} \text{ is } \epsilon\text{-super sufficient} && (29)\\
&\leq 2\epsilon && S^*_{\textsf{nec}} \text{ is } \epsilon\text{-necessary} && (30)\\
&\implies \epsilon \geq \frac{\rho(f(\mathbf{x}), f_\emptyset(\mathbf{x}))}{2} && && (31)
\end{aligned}$$

which is a contradiction because $0 \leq \epsilon < \frac{\rho(f(\mathbf{x}), f_\emptyset(\mathbf{x}))}{2}$. Thus $S^*_{\textsf{suf}} \cap S^*_{\textsf{nec}} \neq \emptyset$. The proof of this result assuming $S^*_{\textsf{nec}}$ is ϵ-super necessary follows the same argument. □

A.1.3 Proof of Theorem 4.1

Theorem 4.1. Let $\tau_1, \tau_2 > 0$ and $0 \leq \epsilon < \frac{1}{2} \cdot \rho(f(\mathbf{x}), f_\emptyset(\mathbf{x}))$. Denote $S^*_{\textsf{suf}}$ and $S^*_{\textsf{nec}}$ to be ϵ-super sufficient and ϵ-super necessary solutions to ($\mathrm{P}_{\textsf{suf}}$) and ($\mathrm{P}_{\textsf{nec}}$), respectively, such that $|S^*_{\textsf{suf}}\,| = \tau_1$ and $|S^*_{\textsf{nec}}| = \tau_2$. Then, there exists a set S^* such that

$$\Delta^{\textsf{uni}}_{\mathcal{V}}(S^*, f, \mathbf{x}, \alpha) \leq \epsilon \quad \text{and} \quad \max(\tau_1, \tau_2) \leq |S^*| < \tau_1 + \tau_2. \tag{32}$$

Furthermore, if $S^*_{\textsf{suf}} \subseteq S^*_{\textsf{nec}}$ or $S^*_{\textsf{nec}} \subseteq S^*_{\textsf{suf}}$. then $S^* = S^*_{\textsf{nec}}$ or $S^* = S^*_{\textsf{suf}}$, respectively.

Proof. Consider the set $S^* = S^*_{\textsf{suf}} \cup S^*_{\textsf{nec}}$. This set has the following properties:

(P1) S^* is ϵ-sufficient because $S^*_{\textsf{suf}}$ is ϵ-super sufficient

(P2) S^* is ϵ-necessary because $S^*_{\textsf{suf}}$ is ϵ-super necessary

(P3) $|S^*| \geq \max(\tau_1, \tau_2)$ with $|S^*| = \tau_1$ when $S^*_{\textsf{nec}} \subset S^*_{\textsf{suf}}$ and with $|S^*| = \tau_2$ when $S^*_{\textsf{suf}} \subset S^*_{\textsf{nec}}$

(P4) Via Lemma 4.1, we know $S^*_{\textsf{suf}} \cap S^*_{\textsf{nec}} \neq \emptyset$ thus $|S^*| < \tau_1 + \tau_2$

Then by (P1) and (P2)

$$\begin{aligned}
\Delta^{\textsf{uni}}_{\mathcal{V}}(S^*, f, \mathbf{x}, \alpha) &= \alpha \cdot \Delta^{\textsf{suf}}_{\mathcal{V}}(S^*, f, \mathbf{x}) + (1-\alpha) \cdot \Delta^{\textsf{nec}}_{\mathcal{V}}(S^*, f, \mathbf{x}) && (33)\\
&\leq \alpha \cdot \epsilon + (1-\alpha) \cdot \epsilon = \epsilon && (34)
\end{aligned}$$

and by (P3) and (P4) we have $\max(\tau_1, \tau_2) \leq |S^*| < \tau_1 + \tau_2$, □

A.1.4 Proof of Corollary 5.1

Corollary 5.1. Suppose for any $S \subseteq [d]$, $\mathcal{V}_S = p(\mathbf{X}_S \mid \mathbf{X}_{S^c} = \mathbf{x}_{S^c})$. Let $\alpha \in (0, 1)$, $\epsilon \geq 0$, and denote $\rho : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ to be a metric on $\mathbb{R}$. Furthermore, for $f(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$ and $\tau > 0$, let S^* be a solution to ($\mathrm{P}_{\textsf{uni}}$) such that $\Delta^{\textsf{uni}}_{\mathcal{V}}(S, f, \mathbf{x}, \alpha) = \epsilon$. Then, S^* satisfies the following conditional independence relations,

$$\rho\left(\mathbb{E}[Y \mid \mathbf{x}],\ \mathbb{E}[Y \mid \mathbf{X}_{S^*} = \mathbf{x}_{S^*}]\right) \leq \frac{\epsilon}{\alpha} \quad \text{and} \quad \rho\left(\mathbb{E}[Y \mid \mathbf{X}_{S^*_c} = \mathbf{x}_{S^*_c}],\ \mathbb{E}[Y]\right) \leq \frac{\epsilon}{1-\alpha}. \tag{35}$$

Proof. All we need to show is that when $\mathcal{V}_S = p(\mathbf{X}_S \mid \mathbf{X}_{S^c} = \mathbf{x}_{S^c})$ and $f(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$, we have

$$f_S(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X}_S = \mathbf{x}_S]. \tag{36}$$

Once this is proven, we can simply apply Lemma 4.1.

To this end, we have by assumption that $f(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ and, for any $S \subseteq [d]$, $\mathcal{V}_S = p(\mathbf{X}_S \mid \mathbf{X}_{S^c} = \mathbf{x}_{S^c})$. Then by definition

$$\begin{aligned}
f_S(\mathbf{x}) = \mathbb{E}_{\mathcal{V}_{S^c}}[f(\mathbf{x}_S, \mathbf{X}_{S^c})] &= \int_{\mathcal{X}} f(\mathbf{x}_S, \mathbf{X}_{S^c}) \cdot p(\mathbf{X}_{S^c} \mid \mathbf{X}_S = \mathbf{x}_S)\, d\mathbf{X}_{S^c} && (37)\\
&= \int_{\mathcal{X}} \mathbb{E}[Y \mid \mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{S^c}] \cdot p(\mathbf{X}_{S^c} \mid \mathbf{X}_S = \mathbf{x}_S)\, d\mathbf{X}_{S^c} && (38)\\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} y \cdot p(y \mid \mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{S^c})\, dy \right) \cdot p(\mathbf{X}_{S^c} \mid \mathbf{X}_S = \mathbf{x}_S)\, d\mathbf{X}_{S^c} && (39)\\
&= \int_{\mathcal{Y}} y \left(\int_{\mathcal{X}} p(y, \mathbf{X}_{S^c} \mid \mathbf{X}_S = \mathbf{x}_S)\, d\mathbf{X}_{S^c} \right) dy && (40)\\
&= \int_{\mathcal{Y}} y \cdot p(y \mid \mathbf{X}_S = \mathbf{x}_S)\, dy && (41)\\
&= \mathbb{E}[Y \mid \mathbf{X}_S = \mathbf{x}_S]. && (42)
\end{aligned}$$

By applying Lemma 4.1, we have the desired result. □

A.1.5 PROOF OF THEOREM 5.1

Theorem 5.1. Consider an input $\mathbf{x}$ for which $f(\mathbf{x}) \neq f_\emptyset(\mathbf{x})$. Denote by $\Lambda_d = \{S, S^c\}$ the partition of $[d] = \{1, 2, \ldots, d\}$, and define the characteristic function to be $v(S) = -\rho(f(\mathbf{x}), f_S(\mathbf{x}))$. Then,

$$\phi_S^{\mathsf{shap}}(\Lambda_d, v) \geq \rho(f(\mathbf{x}), f_\emptyset(\mathbf{x})) - \Delta_{\mathcal{V}}^{\mathsf{uni}}(S, f, \mathbf{x}, \alpha). \tag{43}$$

Proof. Before we prove the result, recall the following properties of a metric ρ in the reals:

(P1) $\forall a, b \in \mathbb{R},\ \rho(a, b) = 0 \iff a = b$

(P2) for $a, b, c \in \mathbb{R}, \quad \rho(a, c) \leq \rho(a, b) + \rho(b, c)$.

Now, for the partition $\Lambda_d = \{S, S^c\}$ of $[d] = \{1, 2, \ldots, d\}$ and characteristic function $v(S) = -\rho(f(\mathbf{x}), f_S(\mathbf{x}))$, $\phi_S^{\mathsf{shap}}(\Lambda_d, v)$ is defined as

$$\begin{aligned}
\phi_S^{\mathsf{shap}}(\Lambda_d, v) &= \frac{1}{2} \cdot [v(S \cup S^c) - v(S^c)] + \frac{1}{2} \cdot [v(S) - v(\emptyset)] && (44)\\
&= \frac{1}{2} \cdot [\rho(f(\mathbf{x}), f_{S^c}(\mathbf{x})) - \rho(f(\mathbf{x}), f(\mathbf{x}))] + \frac{1}{2} \cdot [\rho(f(\mathbf{x}), f_\emptyset(\mathbf{x})) - \rho(f(\mathbf{x}), f_S(\mathbf{x}))] && (45)\\
&= \frac{1}{2} \cdot [\rho(f(\mathbf{x}), f_{S^c}(\mathbf{x}))] + \frac{1}{2} \cdot [\rho(f(\mathbf{x}), f_\emptyset(\mathbf{x})) - \rho(f(\mathbf{x}), f_S(\mathbf{x}))] \qquad \text{by (P1)} && (46)
\end{aligned}$$

By (P2)

$$\begin{aligned}
&\rho(f(\mathbf{x}), f_\emptyset(\mathbf{x})) \leq \rho(f(\mathbf{x}), f_{S^c}(\mathbf{x})) + \rho(f_{S^c}(\mathbf{x}), f_\emptyset(\mathbf{x})) && (47)\\
&\qquad \implies \rho(f(\mathbf{x}), f_{S^c}(\mathbf{x})) \geq \rho(f(\mathbf{x}), f_\emptyset(\mathbf{x})) - \rho(f_{S^c}(\mathbf{x}), f_\emptyset(\mathbf{x})). && (48)
\end{aligned}$$

Thus

$$\begin{aligned}
\phi_S^{\mathsf{shap}}(\Lambda_d, v) &= \frac{1}{2} \cdot [\rho(f(\mathbf{x}), f_{S^c}(\mathbf{x}))] + \frac{1}{2} \cdot [\rho(f(\mathbf{x}), f_\emptyset(\mathbf{x})) - \rho(f(\mathbf{x}), f_S(\mathbf{x}))] && (49)\\
&\geq \frac{1}{2} \cdot [\rho(f(\mathbf{x}), f_\emptyset(\mathbf{x})) - \rho(f_{S^c}(\mathbf{x}), f_\emptyset(\mathbf{x}))] + \frac{1}{2} \cdot [\rho(f(\mathbf{x}), f_\emptyset(\mathbf{x})) - \rho(f(\mathbf{x}), f_S(\mathbf{x}))] && (50)\\
&= \rho(f(\mathbf{x}), f_\emptyset(\mathbf{x})) - \Delta_{\mathcal{V}}^{\mathsf{uni}}(S, f, \mathbf{x}, \alpha). && (51)
\end{aligned}$$

□

A.2 Additional Experimental Details

In this section, we include further experimental details. All experiments were performed on a private cluster with 8 NVIDIA RTX A5000 with 24 GB of memory. All scripts were run on PyTorch `2.0.1`, Python `3.11.5`, and CUDA `12.2`.

A.2.1 RSNA CT Hemorrhage

Dataset Details. The RSNA 2019 Brain CT Hemorrhage Challenge dataset (Flanders et al., 2020), contains 752803 images labeled by a panel of board-certified radiologists with the types of hemorrhage present (epidural, intraparenchymal, intraventricular, subarachnoid, subdural).

Implementation. Recall for this experiment, to identify sufficient and necessary masks S for a sample $\mathbf{x}$, we considered the relaxed optimization problem (Fong et al., 2019; Kolek et al., 2022)

$$\underset{S\subseteq[0,1]^d}{\arg\min}\ \Delta_{\mathcal{V}}^{\mathsf{uni}}(S, f, \mathbf{x}, \alpha) + \lambda_1 \cdot ||S||_1 + \lambda_{\text{TV}} \cdot ||S||_{TV}. \tag{52}$$

where $||S||_1$ and $||S||_{TV}$ are the L^1 and Total Variation norm of S, which promote sparsity and smoothness respectively and λ_{Sp} and λ_{Sm} are the associated. To solve this problem, a mask $S \in [0,1]^{512\times 512}$ is initialized with entries $S_i \sim \mathcal{N}(0.5, \frac{1}{36})$. For 1000 iterations, the mask S is iteratively updated to minimize

$$\alpha \cdot |f(\mathbf{x}) - f_S(\mathbf{x})| + (1-\alpha) \cdot |f(\mathbf{x}) - f_S(\mathbf{x})| + \lambda_1 \cdot ||S||_1 + \lambda_{\text{TV}} \cdot ||S||_{TV} \tag{53}$$

where for any S,

$$f_S(\mathbf{x}) = \frac{1}{K}\sum_{i=1}^{K} f((\tilde{\mathbf{X}}_S)_i) \quad \text{with} \quad (\tilde{\mathbf{X}}_S)_i = \mathbf{x} \circ \tilde{\mathbb{1}}_S + (1 - \tilde{\mathbb{1}}_S) \circ b_i. \tag{54}$$

Here the entries $(\tilde{\mathbb{1}}_S)_i \sim \text{Bernoulli}(S_i)$ and b_i is the ith entry of a vector $\mathbf{b} = (b_1, \cdots, b_d) \sim \mathcal{V}$. In our implementation the reference distribution $\mathcal{V}$ is the unconditional mean image over the of training images and so b_i is the simply the average value of the ith pixel over the training set. To allow for differentiation during optimization, we generate discrete samples $\tilde{\mathbb{1}}_S$ using the Gumbel-Softmax distribution. This methodology simply implies the entries $(\tilde{\mathbf{X}}_S)_i$ is a Bernoulli distribution with outcomes $\{b_i, x_i\}$, i.e. $(\tilde{\mathbf{X}}_S)_i$ is distributed as

$$\Pr[(\tilde{\mathbf{X}}_S)_i = x_i] = S_i \tag{55}$$

$$\Pr[(\tilde{\mathbf{X}}_S)_i = b_i] = 1 - S_i \tag{56}$$

For each $\alpha \in \{0, 0.5, 1\}$, during optimization we set $K = 10$, $\lambda_1 = 3$ and $\lambda_{\text{TV}} = 20$ and use the Adam optimizer with default β-parameters of $\beta_1 = 0.9$, $\beta_2 = 0.99$ and a fixed learning rate of 0.01.

A.2.2 CELEBA-HQ

Dataset Details. We use a modified version of the CelebA-HQ dataset (Lee et al., 2020; Karras, 2017) which contains 30,000 celebrity faces resized to 256×256 pixels with several landmark locations and binary attributes (e.g., eyeglasses, bangs, smiling).

Implementation. Recall for this experiment, to generate sufficient or necessary masks S for samples $\mathbf{x}$, we learn a model $g_\theta : \mathcal{X} \mapsto [0,1]^d$ via solving the following optimization problem:

$$\arg\min_{\theta \in \Theta} \mathop{\mathbb{E}}_{\mathbf{X} \sim \mathcal{D}_\mathcal{X}} \left[\Delta_\mathcal{V}^{\mathsf{uni}}(g_\theta(\mathbf{X}), f, \mathbf{X}, \alpha) + \lambda_1 \cdot ||g_\theta(\mathbf{X})||_1 + \lambda_{\text{TV}} \cdot ||g_\theta(\mathbf{X})||_{\text{TV}} \right] \tag{57}$$

To learn sufficient and necessary explainer models, we solve Eq. (8) via empirical risk minimization for $\alpha \in \{0,1\}$ respectively. Given N samples $\{\mathbf{X}_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mathcal{D}_X$, we solve

$$\frac{1}{N} \sum_{i=1}^{N} \left[\Delta_\mathcal{V}^{\mathsf{uni}}(g_\theta(\mathbf{X}_i), f, \mathbf{X}_i, \alpha) + \lambda_1 \cdot ||g_\theta(\mathbf{X}_i)||_1 + \lambda_{\text{TV}} \cdot ||g_\theta(\mathbf{X}_i)||_{\text{TV}} \right]. \tag{58}$$

Here

$$\Delta_\mathcal{V}^{\mathsf{uni}}(g_\theta(\mathbf{x}_i), f, \mathbf{x}_i, \alpha) = \alpha \cdot |f(\mathbf{x}_i) - f_S(\mathbf{x}_i)| + (1-\alpha) \cdot |f(\mathbf{x}_i) - f_S(\mathbf{x}_i)| \tag{59}$$

where is $f_S(\mathbf{x}_i)$ is evaluated in the same manner as in the RSNA experiment. For $\alpha = 0$, $\lambda_1 = 0.1$ and $\lambda_{\text{TV}} = 100$. For $\alpha = 1$, $\lambda_1 = 1$ and $\lambda_{\text{TV}} = 10$. For both α, during optimization we use a batch size of 32, set $K = 10$ and use the Adam optimizer with default β-parameters of $\beta_1 = 0.9$, $\beta_2 = 0.99$ and a fixed learning rate of 1×10^{-4}

Sampling. To generate the samples in Figs. 4 and 5, samples we use the `CoPaint` method (Zhang et al., 2023). We utilize their code base and pretrained diffusion models with the exact the same parameters as reported in the paper to perform conditional generation. Everything used is available at https://github.com/UCSB-NLP-Chang/CoPaint.

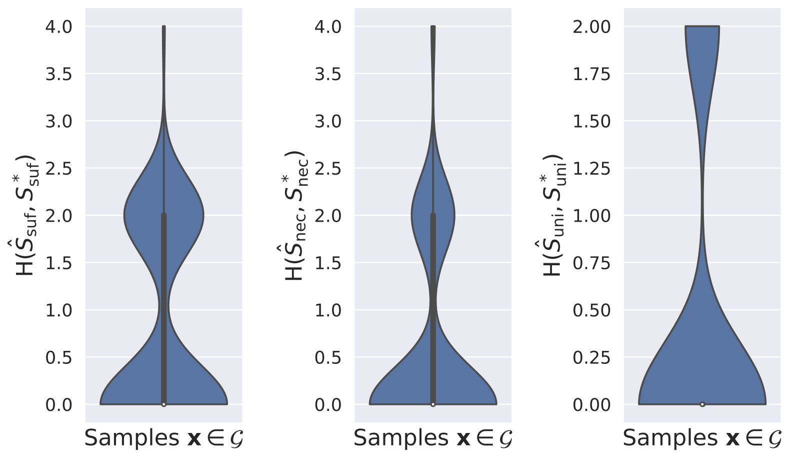


Figure 7: Hamming distances between computed and optimal solutions for P_{suf}, P_{nec}, and P_{uni}

A.3 Additional Experiments

A.3.1 Synthetic Example

We model features $\mathbf{X} \in \mathbb{R}^7$, where $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i \in \{1, 4, 5, 6, 7\}$. The remaining features and response Y follow:

$$X_2 = 2 \cdot X_1 + \epsilon, \quad Y = 4 \cdot X_2 \cdot \mathbf{1}_{\{X_2 > 10\}} + \epsilon, \quad X_3 = 4 \cdot Y + 15 \cdot X_4 \cdot \mathbf{1}_{\{X_4 > 0.5\}} + \epsilon \tag{60}$$

where $\epsilon \sim \mathcal{N}(0, 1)$. For $\mathbf{X} \in \mathcal{G} := \{\mathbf{X} \mid X_2 > 10, \ X_4 > \frac{1}{2}\}$, the data-generating process is represented by the directed acyclic graph (DAG) shown in Fig. 6 (note X_5, X_6 and X_7 are not depicted since they share no dependencies with any of the random variables). We can see that $Y \perp\!\!\!\perp \mathbf{X}_{\{1,5,6,7\}} | \mathbf{X}_{2,3,4}$ and $Y \perp\!\!\!\perp \mathbf{X}_{\{4,5,6,7\}}$. Thus, for $f(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$ and $\mathcal{V}_S = p(\mathbf{X}_{S^c} \mid \mathbf{x}_S)$, the solutions to P_{suf}, P_{nec}, and P_{uni} with $\tau = 4$ are:

$$S^*_{\text{suf}} = \{2, 3, 4\}, \quad S^*_{\text{nec}} = \{1, 2, 3\}, \quad S^*_{\text{uni}} = \{1, 2, 3, 4\}.$$

In this experiment, we train a general predictor (a three-layer fully-connected neural network) to approximate $\mathbb{E}[Y \mid \mathbf{X}]$ and

1. Validate the sets listed above are the optimal solutions.
2. Demonstrate that common post-hoc interpretability methods struggle to recover these solutions.

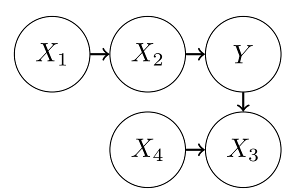
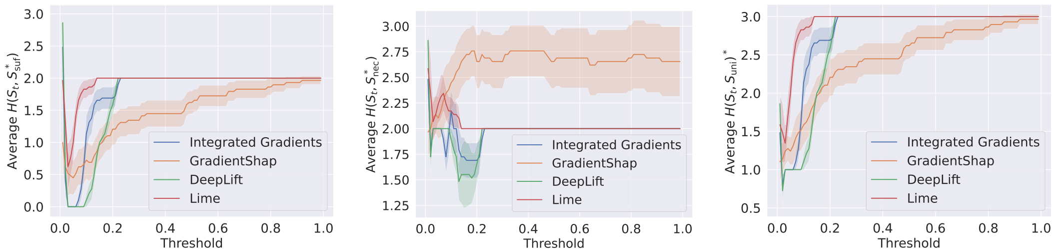


Figure 6: DAG modeling the data-generating process for $\mathbf{X} \in \mathcal{G}$

Validation of Solutions. For $\text{type} \in \{\text{suf}, \text{nec}, \text{uni}\}$, $\tau = 4$, and 100 samples $\mathbf{x} \in \mathcal{G}$ we compute solutions to P_{type}, denoted as $\hat{S}_{\text{type}}$, via exhaustive search. Fig. 7 shows that for all three problems, the Hamming distance between $\hat{S}_{\text{type}}$ and S^*_{ptype} is equal to 0 for a majority of the samples in $\mathcal{G}$. These results indicate that the solutions computed via an exhaustive search do typically retrieve the correct solutions (the minor discrepancies are due to $f(\mathbf{X})$ being an approximation of $\mathbb{E}[Y \mid \mathbf{X}]$). More importantly, this setting is a clear example of how the unified approach provides a different perspective of importance. One would not be able to identify the set $S = \{1, 2, 3, 4\}$ as the most important one without directly solving the unified problem.

Comparison with Post-hoc Methods For our model f and samples $\mathbf{x} \in \mathcal{G}$, we use Integrated Gradients, Gradient Shapley, Deeplift, and Lime to generate attribution scores. To identify whether these methods highlight sufficient and/or necessary features, and as done with our other experiments, we perform the following steps on the attribution scores for a sample $\mathbf{x}$ (so that the outputs of all methods are comparable)

1. We normalize the scores to the interval $[0, 1]$ via min/max normalization.



(a) Hamming distance to S^*_{suf} (b) Hamming distance to S^*_{nec} (c) Hamming distance to S^*_{uni}

Figure 8: Comparison of various post-hoc methods

2. We generate binary masks S_t by thresholding the normalized scores with thresholds $t \in (0, 1)$
3. For $\mathsf{type} \in \{\mathsf{suf}, \mathsf{nec}, \mathsf{uni}\}$, we compute $H(S_t, S^*_{\mathsf{type}})$, the Hamming distance between S_t and the true solutions to P_{suf}, P_{nec}, and P_{uni}

The results in Fig. 8 illustrate that, in general, current post-hoc methods fail to recover the optimal sufficient, necessary, or unified solutions. For thresholds $t \in [0, 0.1]$, we see that Integrated Gradients and Deeplift recover solutions S_t that match the optimal sufficient solution S^*_{suf}. This indicates these methods are capable of highlighting the sufficient features. Besides this observation, we see that for thresholds $t > 0.2$ and all three problems, nearly all methods recover solutions S_t that have a Hamming distance ≥ 2 to the optimal solution indicating that the solutions S_t and optimal solutions S^* differ by at least two elements. As a result, the conclusion is that most common methods do *not* detect sufficient solutions and *no* methods detect necessary or unified solutions.

A.4 Additional Figures

A.4.1 RSNA CT Hemorrhage

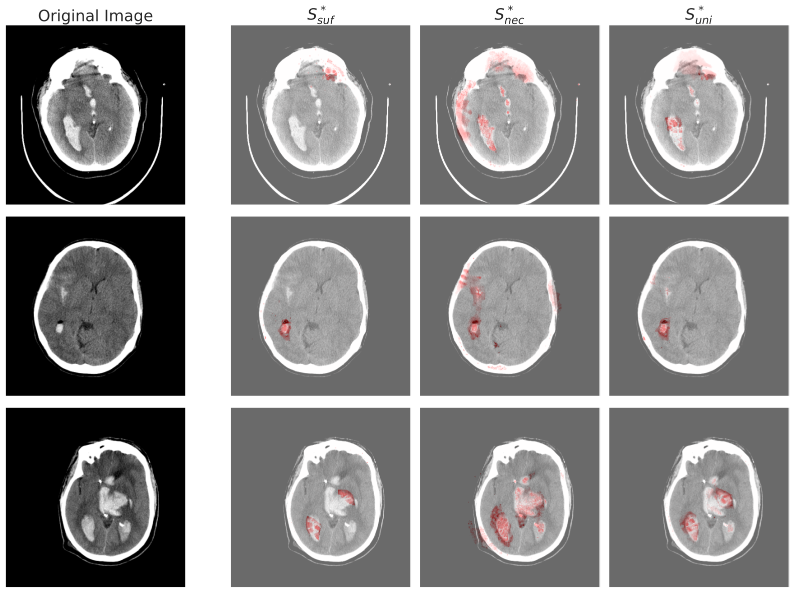


Figure 9: S^*_{suf}, S^*_{nec} and S^*_{uni} for various CT scans.

A.4.2 CelebA-HQ

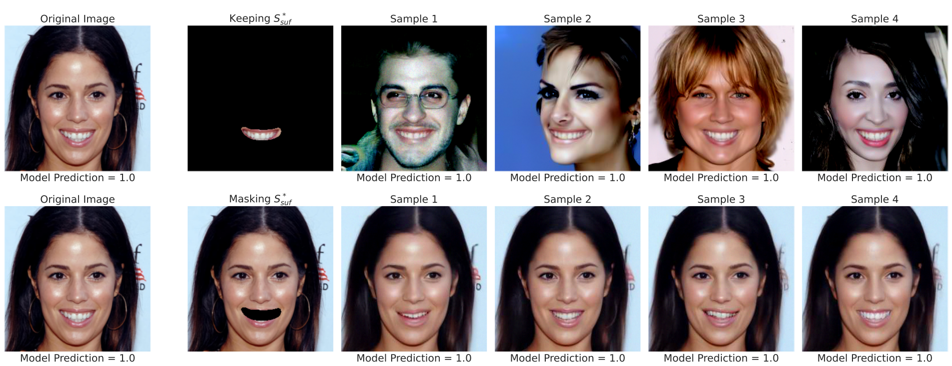


Figure 10: Images and model predictions by fixing and masking the sufficient subset S^*_{suf}

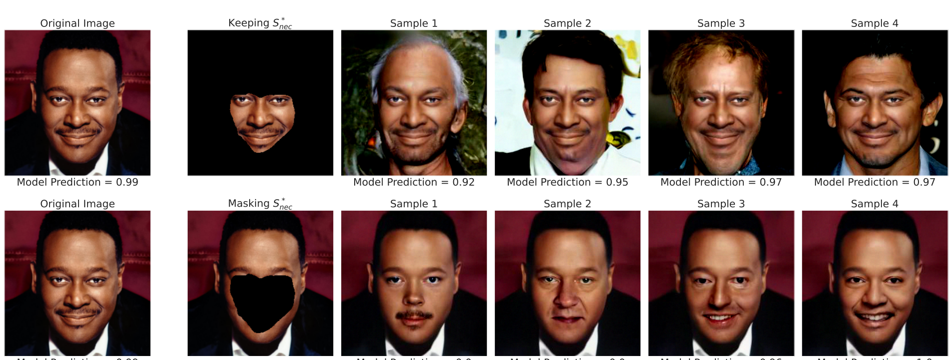


Figure 11: Images and model predictions by fixing and masking the necessary subset S^*_{nec}