Axiomatic Aggregations of Abductive Explanations

Gagan Biradar[1], Yacine Izza[2], Elita Lobo[1], Vignesh Viswanathan[1], Yair Zick[1]

[1]University of Massachusetts, Amherst, USA
[2]CREATE, National University of Singapore, Singapore
{gbiradar,elobo,vviswanathan,yzick}@umass.edu, izza@comp.nus.edu.sg

Abstract

The recent criticisms of the robustness of post hoc model approximation explanation methods (like LIME and SHAP) have led to the rise of model-precise abductive explanations. For each data point, abductive explanations provide a minimal subset of features that are sufficient to generate the outcome. While theoretically sound and rigorous, abductive explanations suffer from a major issue — there can be several valid abductive explanations for the same data point. In such cases, providing a single abductive explanation can be insufficient; on the other hand, providing all valid abductive explanations can be incomprehensible due to their size. In this work, we solve this issue by aggregating the many possible abductive explanations into feature importance scores. We propose three aggregation methods: two based on power indices from cooperative game theory and a third based on a well-known measure of causal strength. We characterize these three methods axiomatically, showing that each of them uniquely satisfies a set of desirable properties. We also evaluate them on multiple datasets and show that these explanations are robust to the attacks that fool SHAP and LIME.

Introduction

The increasing use of complex machine learning (predictive) models in high-stake domains like finance (Ozbayoglu, Gudelek, and Sezer 2020) and healthcare (Pandey et al. 2022; Qayyum et al. 2021) necessitates the design of methods to accurately explain the decisions of these models. Many such methods have been proposed by the AI community. Most of these methods (like SHAP (Lundberg and Lee 2017) and LIME (Ribeiro, Singh, and Guestrin 2016)) explain model decisions by sampling points and evaluating model behavior around a point of interest. While useful in many settings, this class of model approximation-based methods has faced criticisms for being unable to fully capture model behavior (Rudin 2019; Huang and Marques-Silva 2023) and being easily manipulable (Slack et al. 2020). The main issue with these methods stems from the fact that model approximation-based explanation measures use the model's output on a small fraction of the possible input points. This has led to the rise of model-precise *abductive explanations* (AXp's) (Shih, Choi, and Darwiche 2018; Ignatiev, Narodytska, and Marques-Silva 2019) which use the underlying model's structure to compute rigorous explanations. AXp's are simple: they provide a minimal set of features that are sufficient to generate the outcome. In other words, a set of features S forms an AXp for a particular point of interest $\vec{x}$ if no matter how we modify the values of the features outside S, the outcome will not change.

Despite being simple, concise, and theoretically sound, AXp's suffer from a major flaw — there may be several possible AXp's for a given data point. Consider the following example:

Example 1. Suppose that we train a simple rule-based model f for algorithmic loan approval, using the features 'Age', 'Purpose', 'Credit Score', and 'Bank Balance'. The rule-based model has the following closed-form expression:

$$f(\vec{x}) = (\text{Age} > 20 \wedge \text{Purpose} = \text{Education}) \vee (\text{Credit} > 700) \vee (\text{Bank} > 50000)$$

In simple words, if the applicant has an age greater than 20 and is applying for education purposes, the loan is accepted; otherwise, if the applicant has a credit score greater than 700 or a bank account balance greater than 50000, the loan is accepted.

Consider a user with the following details $\vec{x} = (\text{Age} = 30, \text{Purpose} = \text{Education}, \text{Credit} = 750, \text{Bank} = 60000)$. There are three explanations for this point: (Age = 30, Purpose = Education), (Credit = 750), and (Bank = 60000).

In this example, if we provide the AXp (Age = 30, Purpose = Education) to the user, they can infer that their age and purpose played a big role in their decision. However, note that it would be incorrect to infer anything else. The user cannot even tell if the features which are absent from the explanation played any role in their acceptance. In fact, the user still does not know whether the feature Age (present in the explanation) was more important than the feature Credit Score (absent in the explanation). Arguably, Credit Score is more relevant than Age since it is present in a singleton AXp. However, no user presented with only one AXp can make this conclusion.

We propose to aggregate AXp's into importance scores for each feature. Feature importance scores are an extremely well-studied class of explanations (Barocas, Selbst, and Raghavan 2020). As seen with the widespread use of measures like SHAP and LIME, the simple structure of feature

importance scores make it easy to understand and visualize. We propose to use these feature importance scores to give users a comprehensive understanding of model behavior that is impossible to obtain from a single AXp.

Our Contributions

Conceptual. We present three aggregation measures — the Responsibility Index, the Deegan-Packel Index, and the Holler-Packel Index (Section). The Responsibility index is based on the degree of responsibility — a well-known causal strength quantification metric (Chockler, Halpern, and Kupferman 2008). The Deegan-Packel and Holler-Packel indices are based on power indices from the cooperative game theory literature (Deegan and Packel 1978; Holler 1982; Holler and Packel 1983).

Theoretical. For each of these measures, we present an axiomatic characterization, in line with theoretical results in the model explainability community (Patel, Strobel, and Zick 2021; Lundberg and Lee 2017; Datta, Sen, and Zick 2016; Sundararajan and Najmi 2020). Since we deal with aggregating abductive explanations as opposed to conventional model outputs, our proof styles and axioms are novel.

Empirical. We empirically evaluate our measures, comparing them with well-known feature importance measures: SHAP (Lundberg and Lee 2017) and LIME (Ribeiro, Singh, and Guestrin 2016). Our experimental results demonstrate the robustness of our methods, showing specifically that they are capable of identifying biases in a model that SHAP and LIME cannot identify.

Related Work

Abductive explanations were first formally defined in Ignatiev, Narodytska, and Marques-Silva (2019) as a generalization of prime implicant explanations defined in Shih, Choi, and Darwiche (2018). For most commonly used machine learning models models, computing abductive explanations is an intractable problem (Marques-Silva and Ignatiev 2022); hence, computing abductive explanations for these models often requires using NP oracles (e.g. SAT/SMT, MILP, etc).

These oracles have been used in different ways to compute abductive explanations for different classes of models. For example, MILP-encodings have been used for neural networks (Ignatiev, Narodytska, and Marques-Silva 2019) and SMT-encodings have been used for tree ensembles (Ignatiev et al. 2022). For less complex models such as monotonic classifiers and naive bayes classifiers, polynomial time algorithms to compute abductive explanations are known (Marques-Silva et al. 2020, 2021).

The main focus of these papers has been the runtime of the proposed algorithms. There are fewer papers analysing the quality of the output abductive explanations. Notably, the work of Audemard et al. (2022) is also motivated by the fact that there can be several abductive explanations for a single data point; however, their solution is radically different from ours. They propose to use the explainer's preferences over the set of explanations to find a *preferred* abductive explanation to provide to the user.

Huang and Marques-Silva (2023) observe that SHAP (Lundberg and Lee 2017) often fails to identify features that are irrelevant to the prediction of a data point, i.e. assigns a positive score to features that never appear in any abductive explanations. They propose aggregating abductive explanations as an alternative to SHAP but do not propose any concrete measures to do so. Our work answers this call with three axiomatically justified aggregation measures. Parallel to our work[1], Yu, Ignatiev, and Stuckey (2023) also build on the observations of Huang and Marques-Silva (2023) and develop a MARCO-like method (Liffiton et al. 2016) for computing feature importance explanations by aggregating abductive explanations. Their work proposes two aggregation measures, *formal feature attribution (ffa)* and *weighted ffa*, that correspond exactly to the Holler-Packel and Deegan-Packel indices respectively. The results shown in Yu, Ignatiev, and Stuckey (2023), demonstrate empirically the efficiency of the proposed solution. We remark, however, that their work does not offer an axiomatic characterization of these measures, and focuses solely on empirical performance.

There has also been recent work generalizing abductive explanations to *probabilistic abductive explanations* (Wäldchen et al. 2021; Arenas et al. 2022; Izza et al. 2023). Probabilistic abductive explanations allow users to trade-off precision for size, resulting in smaller explanations with lower precision i.e. smaller explanations which are not as robust as abductive explanations.

Our work also contributes novel feature importance measures. Feature importance measures have been well studied in the literature with measures like SHAP (Lundberg and Lee 2017) and LIME (Ribeiro, Singh, and Guestrin 2016) gaining significant popularity. There are several other measures in the literature, many offering variants of the Shapley value (Sundararajan and Najmi 2020; Frye, Rowat, and Feige 2020; Sundararajan, Taly, and Yan 2017). Other works use the Banzhaf index (Patel, Strobel, and Zick 2021) and necessity and sufficiency scores (Galhotra, Pradhan, and Salimi 2021; Watson et al. 2021).

Preliminaries

We denote vectors by $\vec{x}$ and $\vec{y}$. We denote the i-th and j-th indices of the vector $\vec{x}$ using x_i and x_j. Given a set S, we denote the restricted vector containing only the indices $i \in S$ using $\vec{x}_S$. We also use $[k]$ to denote the set $\{1, 2, \ldots, k\}$.

We have a set of features $N = \{1, 2, \ldots, n\}$, where each $i \in N$ has a domain $\mathcal{X}_i$. We use $\mathcal{X} = \times_{i \in N} \mathcal{X}_i$ to denote the domain of the input space. We are given a *model of interest* $f \in \mathcal{F}$ that maps input vectors $\vec{x} \in \mathcal{X}$ to a binary output variable $y \in \{0, 1\}$. In the local post hoc explanation problem, we would like to explain the output of the model of interest f on a *point of interest* $\vec{x}$. We work with two forms of model explanations in this paper.

The first is that of *feature importance weights* (or *feature importance scores*): feature importance weights provide a score to each feature proportional to their importance in the

[1]The work of Yu, Ignatiev, and Stuckey (2023) was developed independently and in parallel to our work.

generation of the outcome $f(\vec{x})$. Commonly used feature importance measures are LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017).

Second, an *abductive explanation* (AXp) for a point of interest $\vec{x}$ is a minimal subset of features which are sufficient to generate the outcome $f(\vec{x})$. More formally, an AXp (as defined by Ignatiev et al. (2022)) corresponds to a subset minimal set of features S such that

$$\forall \vec{y} \in \mathcal{X},\ \big(\vec{y}_S = \vec{x}_S\big) \implies \big(f(\vec{y}) = f(\vec{x})\big). \qquad (1)$$

By subset minimality, if S satisfies (1), then no proper subset of S satisfies (1). We use $\mathcal{M}(\vec{x}, f)$ to denote the set of AXp's for a point of interest $\vec{x}$ under a model of interest f. We also use $\mathcal{M}_i(\vec{x}, f)$ to denote the subset of $\mathcal{M}(\vec{x}, f)$ containing all the AXp's with the feature i. Our goal is to create aggregation measures that maps $\mathcal{M}(\vec{x}, f)$ to an importance score for each feature $i \in N$.

A Cooperative Game Theory Perspective

In this paper, we propose to aggregate abductive explanations into feature importance scores[2].

A common approach used to compute feature importance scores is via modeling the problem as a cooperative game (Patel, Strobel, and Zick 2021; Datta, Sen, and Zick 2016; Lundberg and Lee 2017). This formulation allows us to both, tap into the existing literature on power indices (like the Shapley value) to create feature importance measures, as well as use theoretical techniques from the literature to provide axiomatic characterizations for new measures. In this paper, we do both.

A simple cooperative game (Chalkiadakis, Elkind, and Wooldridge 2011) (N, v) is defined over a set of players N and a monotone[3] binary value function $v : 2^N \mapsto \{0, 1\}$. The set of players, in our setting (and several others (Patel, Strobel, and Zick 2021; Datta, Sen, and Zick 2016; Lundberg and Lee 2017)), are the features of the model of interest N. The value function v loosely represents the value of each (sub)set of players; in model explanations, the value function represents the joint importance of a set of features in generating the outcome.

A set $S \subseteq N$ is referred to as a *minimal winning set* if $v(S) = 1$ and for all proper subsets $T \subset S$, $v(T) = 0$. Minimal winning sets are a natural analog of AXp's in the realm of cooperative game theory. There are specific power indices like the Deegan-Packel index (Deegan and Packel 1978) and the Holler-Packel index (Holler and Packel 1983; Holler 1982) which take as input the set of all minimum winning sets and output a score corresponding to each player (in our case, feature) in the cooperative game. These measures are natural candidates to convert AXp's into feature importance scores.

[2]Feature importance scores computed using AXp's are also referred to as *formal feature attribution* (FFA) in Yu, Ignatiev, and Stuckey (2023).

[3]Recall that a set function v is monotonic if for all $S \subseteq T \subseteq N$, $v(S) \leq v(T)$.

A Framework for Abductive Explanation Aggregation

Formally, we define an *abductive explanation aggregator* (or simply an *aggregator*) as a function that maps a point $\vec{x}$ and a model f to a vector in $\mathbb{R}^n$ using only the AXp's of the point $\vec{x}$ under the model f; the output vector can be interpreted as importance scores for each feature. For any arbitrary aggregator $\beta : \mathcal{X} \times \mathcal{F} \to \mathbb{R}^n$, we use $\beta_i(\vec{x}, f)$ as the importance weight given to the i-th feature for a specific datapoint-model pair $(\vec{x}, f)$.

In order to design meaningful aggregators, we take an axiomatic approach: we start with a set of desirable properties and then find the unique aggregator which satisfies these properties. This is a common approach in explainable machine learning (Datta, Sen, and Zick 2016; Lundberg and Lee 2017; Sundararajan, Taly, and Yan 2017; Patel, Strobel, and Zick 2021) and more recently, boolean functions (Harder et al. 2023). The popular Shapley value (Young 1985) is the unique measure that satisfies four desirable properties — Monotonicity, Symmetry, Null Feature, and Efficiency.

However, the exact definitions of these four properties in the characterization of the Shapley value do not extend to our setting. Moreover, the Shapley value does not aggregate AXp's (or more generally, minimal winning sets). Therefore, for our axiomatic characterization, we formally define variants of these properties, keeping the spirit of these definitions intact. We present these definitions below.

α-Monotonicity: Let α be some function that quantifies the relevance of a set of AXp's a feature i is present in. A feature importance score is monotonic with respect to α if for each feature i and dataset model pair $(\vec{x}, f)$, the importance score given to i is monotonic with respect to $\alpha(\mathcal{M}_i(\vec{x}, f))$.

In simple words, the higher the rank of the set of AXp's containing a feature (according to α), the higher their importance scores. The ranking function α can capture several intuitive desirable properties. For example, if we want features present in a larger number of AXp's to receive a higher score, we can simply set $\alpha(\mathcal{S}) = |\mathcal{S}|$. Otherwise, if we want features present in smaller explanations to receive a higher score, we set $\alpha(\mathcal{S}) = -\min_{S \in \mathcal{S}} |S|$.

Formally, let $\alpha : 2^{2^N} \mapsto \mathcal{Y}$ be a function that ranks sets of AXp's, i.e., maps every set of AXp's to a partially ordered set $\mathcal{Y}$. An aggregator β is said to satisfy α-monotonicity if for any two datapoint-model pairs $(\vec{x}, f)$ and $(\vec{y}, g)$ and a feature i, $\alpha(\mathcal{M}_i(\vec{x}, f)) \leq \alpha(\mathcal{M}_i(\vec{y}, g))$ implies $\beta_i(\vec{x}, f) \leq \beta_i(\vec{y}, g)$. Additionally, if the feature i has the same set of AXp's under $(\vec{x}, f)$ and $(\vec{y}, g)$ — i.e., $\mathcal{M}_i(\vec{x}, f) = \mathcal{M}_i(\vec{y}, g)$ — then $\beta_i(\vec{x}, f) = \beta_i(\vec{y}, g)$.

Symmetry: This property requires that the index of a feature should not affect its score. That is, the score of feature i should not change if we change its position. Given a permutation $\pi : N \to N$, we define $\pi\vec{x}$ as the reordering of the feature values in $\vec{x}$ according to π. In addition, given a permutation $\pi : N \to N$, we define πf as the function that results from permuting the input point using π before computing the output. More formally, $\pi f(\vec{x}) = f(\pi\vec{x})$. We are now ready to formally define the symmetry property:

Measure	α**-Monotonicity**	C**-Efficiency**
Holler-Packel Index $\eta_i(\vec{x}, f) = \lvert\mathcal{M}_i(\vec{x}, f)\rvert$	$\alpha(\mathcal{S}) = \mathcal{S}$ and $\alpha(\mathcal{S}) \leq \alpha(\mathcal{T})$ iff $\mathcal{S} \subseteq \mathcal{T}$	$C(\vec{x}, f) = \sum_{i \in N} \lvert\mathcal{M}_i(\vec{x}, f)\rvert$
Deegan-Packel Index $\phi_i(\vec{x}, f) = \sum_{S \in \mathcal{M}_i(\vec{x}, f)} \frac{1}{\lvert S\rvert}$	$\alpha(\mathcal{S}) = \mathcal{S}$ and $\alpha(\mathcal{S}) \leq \alpha(\mathcal{T})$ iff $\mathcal{S} \subseteq \mathcal{T}$	$C(\vec{x}, f) = \lvert\mathcal{M}(\vec{x}, f)\rvert$
Responsibility Index $\rho_i(\vec{x}, f) = \max_{S \in \mathcal{M}_i(\vec{x}, f)} \frac{1}{\lvert S\rvert}$	$\alpha(\mathcal{S}) = -\min_{S \in \mathcal{S}} \lvert S\rvert$	NA

Table 1: A summary of the α and C values from the Monotonicity and Efficiency properties respectively of each measure defined in this paper. All three measures satisfy Symmetry and Null Feature. The Responsibility index satisfies an alternative efficiency property which is incomparable to C-efficiency.

An aggregator β satisfies symmetry if for any datapoint-model pair $(\vec{x}, f)$ and a permutation π, $\pi\beta(\vec{x}, f) = \beta(\pi\vec{x}, \pi^{-1}f)$.

Null Feature: if a feature is not present in *any* abductive explanation, it is given a score of 0. This property explicitly sets a baseline value for importance scores. More formally, an aggregator η satisfies Null Feature if for any datapoint-model pair $(\vec{x}, f)$ and any feature i, $\mathcal{M}_i(\vec{x}, f) = \emptyset$ implies that $\eta_i(\vec{x}, f) = 0$.

C-Efficiency: This property requires the scores output by aggregators to sum up to a *fixed value*; in other words, for any datapoint-model pair $(\vec{x}, f)$, $\sum_{i \in N} \beta_i(\vec{x}, f)$ must be a fixed value. Not only does efficiency bound the importance scores, but it also ensures that features are not always given a trivial score of 0. The fixed value may depend on the aggregator β, the model f, and the datapoint $\vec{x}$. To capture this, we define a function C that maps each datapoint-model pair $(\vec{x}, f)$ to a real value.

An aggregator β is C-efficient if for any datapoint-model pair $(\vec{x}, f)$, $\sum_{i \in N} \beta_i(\vec{x}, f) = C(\vec{x}, f)$.

We deliberately define the above properties flexibly. There are different reasonable choices of α-monotonicity and C-efficiency — each leading to a different aggregation measure (Table 1). In what follows, we formally present these choices and mathematically find the measures they characterize. It is worth noting, as shown by Huang and Marques-Silva (2023), that the popular SHAP framework fails to satisfy the Null Feature property while all the measures we propose in this paper are guaranteed to satisfy the Null Feature property.

The Holler-Packel Index

We start with the Holler-Packel index, named after the power index in cooperative game theory (Holler 1982; Holler and Packel 1983). The Holler-Packel index measures the importance of each feature as the number of AXp's that contain it. More formally, the Holler-Packel index of a feature i (denoted by $\eta_i(\vec{x}, f)$) is given by

$$\eta_i(\vec{x}, f) = |\mathcal{M}_i(\vec{x}, f)| \tag{2}$$

The Holler-Packel index satisfies a property we call *Minimal Monotonicity*. This property corresponds to α-Monotonicity when $\alpha(\mathcal{S}) = \mathcal{S}$ and $\alpha(S) \leq \alpha(T)$ if and only if $\mathcal{S} \subseteq \mathcal{T}$. Minimal Monotonicity (loosely speaking) ensures that features present in a larger number of AXp's get a higher importance score.

The Holler-Packel index also satisfies C-Efficiency where $C(\vec{x}, f)$ is defined as $\sum_{i \in N} |\mathcal{M}_i(x, f)|$. We refer to this property as $(\sum_{i \in N} |\mathcal{M}_i(x, f)|)$-Efficiency for clarity.

Our first result shows that the Holler-Packel index is the only index that satisfies Minimal Monotonicity, Symmetry, Null Feature, and $(\sum_{i \in N} |\mathcal{M}_i(x, f)|)$-Efficiency.

Theorem 1. *The only aggregator that satisfies Minimal Monotonicity, Symmetry, Null Feature, and $(\sum_{i \in N} |\mathcal{M}_i(x, f)|)$-Efficiency is the Holler-Packel index given by* (2)*.*

The Holler-Packel index was used as a heuristic AXp aggregator in prior work under the term 'hit rate' (Marques-Silva et al. 2020). Theorem 1 theoretically justifies the hit rate.

The Deegan-Packel Index

Next, we present the Deegan-Packel index. This method is also named after the similar game-theoretic power index (Deegan and Packel 1978). The Deegan-Packel index, like the Holler-Packel index, counts the number of AXp's a feature is included in but unlike the Holler-Packel index, each AXp is given a weight inversely proportional to its size. This ensures that smaller AXp's are prioritized over larger AXp's. Formally, the Deegan-Packel index is defined as follows:

$$\phi_i(\vec{x}, f) = \sum_{S \in \mathcal{M}_i(\vec{x}, f)} \frac{1}{|S|} \tag{3}$$

Note that this aggregator also satisfies Minimal Monotonicity, Symmetry, and Null Feature. However, the Deegan-Packel index satisfies a different notion of C-Efficiency. The efficiency notion satisfied by the Deegan-Packel index corresponds to C-Efficiency where $C(\vec{x}, f)$ is defined as $|\mathcal{M}(\vec{x}, f)|$. We refer to this efficiency notion as $|\mathcal{M}(\vec{x}, f)|$-Efficiency for clarity.

Our second result shows that the Deegan-Packel index uniquely satisfies Minimal Monotonicity, Symmetry, Null Feature, and $|\mathcal{M}(\vec{x}, f)|$-Efficiency.

Theorem 2. *The only aggregator that satisfies Minimal Monotonicity, Symmetry, Null Feature, and $|\mathcal{M}(\vec{x}, f)|$-Efficiency is the Deegan-Packel index given by* (3)*.*

The Responsibility Index

We now present our third and final aggregator, the Responsibility index, named after the degree of responsibility (Chockler, Halpern, and Kupferman 2008; Chockler and Halpern 2004) used commonly to measure causal strength.

The Responsibility index (denoted by ρ) of a feature is the inverse of the size of the smallest AXp containing that feature. More formally,

$$\rho_i(\vec{x}, f) = \begin{cases} \max_{S \in \mathcal{M}_i(\vec{x}, f)} \frac{1}{|S|} & \mathcal{M}_i(\vec{x}, f) \neq \emptyset \\ 0 & \mathcal{M}_i(\vec{x}, f) = \emptyset \end{cases} \quad (4)$$

To characterize this aggregator, we require different versions of Monotonicity and Efficiency. Our new monotonicity property requires aggregators to provide a higher score to features present in smaller AXp's. We refer to this property as Minimum Size Monotonicity: this corresponds to α-Monotonicity where given a set of AXp's $\mathcal{S}$, we let $\alpha(\mathcal{S}) = -\min_{S \in \mathcal{S}} |S|$.

The new efficiency property does not fit into the C-Efficiency framework used so far and is easier to define as two new properties — Unit Efficiency and Contraction. Unit Efficiency requires that the score given to any feature present in a singleton AXp be 1. This property is used to upper bound the score given to a feature.

Unit Efficiency: For any datapoint-model pair $(\vec{x}, f)$, $\mathcal{M}_i(\vec{x}, f) = \{\{i\}\}$ implies $\rho_i(\vec{x}, f) = 1$.

To define the contraction property, we define the *contraction operation* on the set of features N: we replace a subset of features $T \subseteq N$ by a single feature $[T]$ corresponding to the set. The *contracted data point* $\vec{x}^{[T]}$ is the same point as $\vec{x}$, but we treat all the features in T as a single feature $[T]$. The contraction property requires that a contracted feature $[T]$ does not receive a score greater than the sum of the scores given to the individual features in T.

Contraction: For any subset T that does not contain a null feature (i.e., a feature not included in any AXp), we have $\rho_{[T]}(\vec{x}^{[T]}, f) \leq \sum_{i \in T} \rho_i(\vec{x}, f)$. Moreover, equality holds if $T \in \{S : S \in \arg\min_{S' \in \mathcal{M}_i(\vec{x}, f)} |S'|\}$ for all $i \in T$. In other words, equality holds iff T is the smallest AXp for all the features in T.

The contraction property bounds the gain one gets by combining features and ensures that the total attribution that a set of features receives when combined does not exceed the sum of the individual attributions of each element in the set.

We are now ready to present our characterization of the Responsibility index.

Theorem 3. *The Responsibility index is the only aggregator which satisfies Minimum Size Monotonicity, Unit Efficiency, Contraction, Symmetry, and Null Feature.*

Impossibilities

The framework discussed above can be used to axiomatically characterize several indices. Our axiomatic approach also offers insights as to what *can* be accomplished by aggregating AXp's. We prove that some choices of α and C may create a set of properties that are impossible to satisfy simultaneously. For example, the Shapley value's efficiency property stipulates that all Shapley values must sum to 1. Somewhat surprisingly, this is not possible when taking an AXp approach.

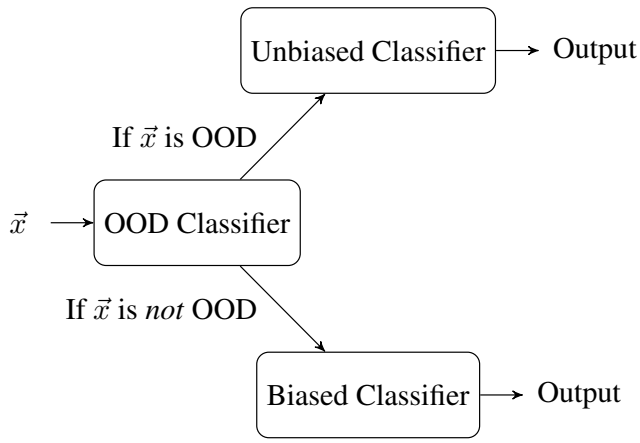

Figure 1: A pictorial description of the attack model. OOD is short for out-of-distribution.

Proposition 4. *There exists no aggregator satisfying Minimal Monotonicity, Symmetry, Null Feature, and 1-Efficiency.*

All the indices described in this section inherit the precision and robustness of AXp's while simultaneously satisfying a set of desirable properties. In what follows, we demonstrate the value of this robustness empirically.

Empirical Evaluation

To showcase the robustness of the explanations generated by our methods, we study their empirical behavior against adversarial attacks proposed by Slack et al. (2020). Specifically, we investigate if our framework successfully uncovers underlying biases in adversarial classifiers that popular explanation methods like LIME and SHAP often fail to identify (Slack et al. 2020). We describe the details of the datasets used in our experiments below.

Compas (Angwin et al. 2016): This dataset contains information about the demographics, criminal records, and Compas risk scores of 6172 individual defendants from Broward County, Florida. Individuals are labeled with either a 'high' or 'low' risk score, with race as the sensitive feature.

German Credit (Dua and Graff 2017): This dataset contains financial and demographic information on 1000 loan applicants. Each candidate is labeled as either a good or bad loan candidate. The sensitive feature is gender.

Attack Model. We evaluate the robustness of our explanation methods using the adversarial attacks proposed by Slack et al. (2020) for LIME and SHAP. The underlying attack model is a two-level adversarial classifier in both adversarial attacks. The first level of the adversarial classifier is an out-of-distribution (OOD) classifier that predicts if a point is OOD or not. The second level of the adversarial classifier consists of a biased and unbiased prediction model, both

Features	Lime (%)			Responsibility (%)			Holler-Packel (%)			Deegan-Packel (%)		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
Race	0.0	0.0	0.0	0.921	0.079	0.0	0.845	0.148	0.007	0.845	0.148	0.007
UC1	0.492	0.508	0.0	0.601	0.399	0.0	0.157	0.843	0.0	0.157	0.843	0.0
UC2	0.508	0.492	0.0	0.601	0.399	0.0	0.157	0.843	0.0	0.157	0.843	0.0

Table 2: This table shows the results of the LIME attack experiment on the Compas dataset. Each row represents the frequency of occurrence of either a sensitive feature (*Race*) or an uncorrelated feature (*UC1*,*UC2*) in the top 3 positions when ranked based on their LIME scores, Responsibility indices, Holler-Packel indices, and Deegan-Packel indices. LIME explanations do not uncover the underlying biases of the attack model, whereas the Responsibility index, Deegan-Packel index, and Holler-Packel index successfully uncover the underlying biases of the attack model in the explanations they generate.

Features	SHAP (%)			Responsibility (%)			Holler-Packel (%)			Deegan-Packel (%)		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
Race	0.416	0.238	0.141	0.946	0.044	0.01	0.867	0.036	0.052	0.867	0.039	0.057
UC1	0.252	0.249	0.172	0.608	0.316	0.067	0.146	0.47	0.215	0.146	0.552	0.138
UC2	0.215	0.249	0.304	0.618	0.297	0.08	0.148	0.466	0.213	0.148	0.554	0.133

Table 3: This table shows the results of the SHAP attack experiment on the Compas dataset. Each row represents the frequency of occurrence of either a sensitive feature (*Race*) or an uncorrelated feature (*UC1*,*UC2*) in the top 3 positions when ranked based on their SHAP scores, Responsibility indices, Holler-Packel indices, and Deegan-Packel indices.

of which predict the target label for a given data point. The biased prediction model makes predictions solely based on the sensitive feature in the dataset. In contrast, the unbiased prediction model makes predictions using features that are uncorrelated with the sensitive feature.

With the Compas dataset, the biased model uses the *Race* feature for predicting the target label. In contrast, the unbiased model uses either one or two uncorrelated synthetic features (*UC1, UC2*) for predicting the label. When two uncorrelated features are used, the label is their values' *XOR*. Similarly, with the German Credit dataset, the unbiased and biased models use the *LoanRateAsPercentOfIncome* feature and *Gender* feature for predicting the target label, respectively.

For a given data point, the adversarial classifier first uses the OOD classifier to predict whether the given data point is OOD data. If the given data point is OOD data, the adversarial classifier uses the unbiased prediction model to predict the target label; else, the adversarial classifier uses the biased classifier to predict the target label (see Figure 1). Most points in the dataset are classified as in-distribution and therefore, the prediction of the attack model for these points will be made solely using the sensitive feature of the dataset. Since the type of explanations generated by popular methods like LIME and SHAP tend to be heavily influenced by the predictions of the classifier model on OOD data points, this attack is designed to hide the underlying biases of the model by ensuring the bias is only applied to in-distribution data points. For each method (SHAP and LIME), Slack et al. (2020) use a different attack model with the same high-level approach described above. Implementation details[4] of each attack are outlined in the extended version of the paper (Biradar et al. 2023).

Experimental Setup. We split a given dataset into train and test datasets in all our experiments. We use the training dataset to train OOD classifiers for the LIME and SHAP attacks and the test dataset to evaluate our methods' robustness. To generate explanations using our proposed AXp aggregators, we must first compute the set of all AXp's for the adversarial classifier model. We do this using the MARCO algorithm (Liffiton et al. 2016). After generating the complete set of AXp's for the adversarial classifier, we compute the feature importance scores using each of our methods — the Holler-Packel index, Deegan-Packel index, and the Responsibility index. We use these feature importance scores as explanations for each point in the test dataset.

We compare our methods with LIME and SHAP, computed using their respective publicly available libraries (Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2016).

Evaluating Robustness to Adversarial LIME and SHAP attacks. For each data point in the test dataset, we rank features based on the feature importance scores given by each explanation method. Note that we allow multiple features to hold the same rank if they have the same importance scores. For each explanation method, we compute the fraction of data points in which the sensitive and uncorrelated features appear in the top three positions. Since most of the points in the test dataset are 'in-distribution' and classified as such by the OOD classifier, any good explanation method should identify that the adversarial classifier makes its prediction largely based on the sensitive feature for most of the points in the test dataset. In other words, the sensitive feature should receive a high importance score.

Table 2 shows the percentage of data points for which the

[4]Code available at https://github.com/elitalobo/aggrxp

Features	Lime (%)			Responsibility (%)			Holler-Packel (%)			Deegan-Packel (%)		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
Gender	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
LR	1.0	0.0	0.0	0.46	0.54	0.0	0.0	0.69	0.31	0.0	0.72	0.28

Table 4: This table shows the results of the LIME attack experiment on the German Credit dataset. Each row represents the frequency of occurrence of either a sensitive feature (*Gender*) or an uncorrelated feature (*LoanRateAsPercentOfIncome*) in the top 3 positions when ranked based on their LIME scores, Responsibility indices, Holler-Packel indices, and Deegan-Packel indices.

Features	SHAP (%)			Responsibility (%)			Holler-Packel (%)			Deegan-Packel (%)		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
Gender	0.0	0.41	0.01	0.93	0.04	0.03	0.87	0.07	0.02	0.87	0.07	0.02
LR	0.96	0.0	0.04	0.55	0.44	0.01	0.17	0.81	0.02	0.17	0.82	0.0

Table 5: This table shows the results of the SHAP attack experiment on the German Credit dataset. Each row represents the frequency of occurrence of either a sensitive feature (*Gender*) or an uncorrelated feature (*LoanRateAsPercentOfIncome*) in the top 3 positions when ranked based on their SHAP scores, Responsibility indices, Holler-Packel indices, and Deegan-Packel indices.

sensitive attribute (i.e., *Race*) and the uncorrelated features (*UC1* and *UC2*) appear in the top three positions when features are ranked using LIME and our methods in the LIME attack experiment on the Compas dataset.

Similarly, Table 3 shows the percentage of data points for which the sensitive attribute (i.e., *Race*) and the uncorrelated features (*UC1* and *UC2*) appear in the top three positions when features are ranked using SHAP and our methods in the SHAP attack experiment for the Compas dataset.

Since the biased classifier is used to predict the label for almost all the test points, we expect the explanations to assign a high feature importance score to the sensitive feature. However, we observe that in the LIME attack experiment, LIME does not always assign high scores to the sensitive feature — *Race* — due to which *Race* does not at all appear in the top three positions when two uncorrelated features are used. The uncorrelated features are incorrectly ranked higher than the sensitive feature. On the other hand, the Responsibility index, the Holler-Packel index, and the Deegan-Packel index assign the highest feature importance scores to *Race*: *Race* appears in the top position for the majority of the instances ($> 84\%$). It is important to note that the instances in which our explanation methods do not assign a high importance score to the *Race* feature are the instances where the OOD classifier classifies test dataset instances as ODD instances.

We observe a similar pattern to LIME in the SHAP attack experiment. In this experiment, AXp aggregators rank *Race* as the most important feature in at least 86% of test data, whereas SHAP ranks *Race* as the most important feature only for 41.6% of the returned explanations.

We see similar results with the German Credit dataset reported in Table 4 and Table 5. In both LIME and SHAP attacks, we observe that the *LoanRateAsPercentOfIncome* feature appears in the top position for most of the delivered explanations. However, the sensitive feature — *Gender* — does not appear in the top position in any instance.

In contrast, the Responsibility Index, the Holler-Packel Index, and the Deegan-Packel Index correctly assign the highest feature importance score to the sensitive feature — *Gender* — for most of the data points; *Gender* appears in the top position in $> 87\%$ of the instances in both the LIME and SHAP attack experiments. Clearly, we can conclude that our AXp aggregators generate more robust and reliable explanations than LIME and SHAP.

Conclusion and Future Work

In this work, we aggregate abductive explanations into feature importance scores. We present three methods that aggregate abductive explanations, showing that each of them uniquely satisfies a set of desirable properties. We also empirically evaluate each of our methods, showing that they are robust to attacks that SHAP and LIME are vulnerable to.

Our focus in this paper has been the axiomatic characterization and comparison of different measures. We believe an empirical comparison of the three methods we propose is also worth exploring in future work. This study is likely to yield insights into the differences in applicability of each of our three methods, further leading to a deeper understanding into how abductive explanations should be aggregated.

Acknowledgments

This research supported in part by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's software used across the country

to predict future criminals. And it's biased against blacks. *ProPublica*.

Arenas, M.; Barceló, P.; Orth, M. A. R.; and Subercaseaux, B. 2022. On Computing Probabilistic Explanations for Decision Trees. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Audemard, G.; Bellart, S.; Bounia, L.; Koriche, F.; Lagniez, J.-M.; and Marquis, P. 2022. On Preferred Abductive Explanations for Decision Trees and Random Forests. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 643–650.

Barocas, S.; Selbst, A. D.; and Raghavan, M. 2020. The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the 3rd ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 80–89.

Biradar, G.; Izza, Y.; Lobo, E.; Viswanathan, V.; and Zick, Y. 2023. Axiomatic Aggregations of Abductive Explanations. *CoRR*, abs/2310.03131.

Chalkiadakis, G.; Elkind, E.; and Wooldridge, M., eds. 2011. *Computational Aspects of Cooperative Game Theory*. Morgan & Claypool Publishers, 1st edition.

Chockler, H.; and Halpern, J. Y. 2004. Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, 22: 93–115.

Chockler, H.; Halpern, J. Y.; and Kupferman, O. 2008. What Causes a System to Satisfy a Specification? *ACM Trans. Comput. Logic*, 9(3).

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. *IEEE Symposium on Security and Privacy*, 598–617.

Deegan, J.; and Packel, E. 1978. A new index of power for simple n-person games. *International Journal of Game Theory*, 7: 113–123.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository.

Frye, C.; Rowat, C.; and Feige, I. 2020. Asymmetric Shapley Values: Incorporating Causal Knowledge into Model-agnostic Explainability. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Galhotra, S.; Pradhan, R.; and Salimi, B. 2021. Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals. arXiv:2103.11972.

Harder, H.; Jantsch, S.; Baier, C.; and Dubslaff, C. 2023. A Unifying Formal Approach to Importance Values in Boolean Functions. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, 2728–2737.

Holler, M. J. 1982. Forming Coalitions and Measuring Voting Power. *Political Studies*, 30: 262–271.

Holler, M. J.; and Packel, E. W. 1983. Power, Luck and the Right Index. *Journal of Economics*, 43: 21–29.

Huang, X.; and Marques-Silva, J. 2023. The Inadequacy of Shapley Values for Explainability. arXiv:2302.08160.

Ignatiev, A.; Izza, Y.; Stuckey, P. J.; and Marques-Silva, J. 2022. Using MaxSAT for Efficient Explanations of Tree Ensembles. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, 3776–3785.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019. Abduction-Based Explanations for Machine Learning Models. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*.

Izza, Y.; Huang, X.; Ignatiev, A.; Narodytska, N.; Cooper, M. C.; and Marques-Silva, J. 2023. On computing probabilistic abductive explanations. *Int. J. Approx. Reason.*, 159.

Liffiton, M. H.; Previti, A.; Malik, A.; and Marques-Silva, J. 2016. Fast, flexible MUS enumeration. *Constraints An Int. J.*, 21(2): 223–250.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS)*, 4768–4777.

Marques-Silva, J.; Gerspacher, T.; Cooper, M. C.; Ignatiev, A.; and Narodytska, N. 2020. Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Marques-Silva, J.; Gerspacher, T.; Cooper, M. C.; Ignatiev, A.; and Narodytska, N. 2021. Explanations for Monotonic Classifiers. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 7469–7479.

Marques-Silva, J.; and Ignatiev, A. 2022. Delivering Trustworthy AI through Formal XAI. In *AAAI*, 12342–12350.

Ozbayoglu, A. M.; Gudelek, M. U.; and Sezer, O. B. 2020. Deep learning for financial applications : A survey. *Applied Soft Computing*, 93: 106384.

Pandey, B.; Kumar Pandey, D.; Pratap Mishra, B.; and Rhmann, W. 2022. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. *Journal of King Saud University - Computer and Information Sciences*, 34: 5083–5099.

Patel, N.; Strobel, M.; and Zick, Y. 2021. High Dimensional Model Explanations: An Axiomatic Approach. In *Proceedings of the 4th ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 401–411. ACM.

Qayyum, A.; Qadir, J.; Bilal, M.; and Al-Fuqaha, A. 2021. Secure and Robust Machine Learning for Healthcare: A Survey. *IEEE Reviews in Biomedical Engineering*, 14: 156–180.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (KDD)*, 1135–1144.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A Symbolic Approach to Explaining Bayesian Network Classifiers. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 5103–5111.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the 3rd AAAI/ACM Conference on Artifical Intelligence, Ethics, and Society (AIES)*.

Sundararajan, M.; and Najmi, A. 2020. The Many Shapley Values for Model Explanation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 9269–9278.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 3319–3328.

Wäldchen, S.; MacDonald, J.; Hauch, S.; and Kutyniok, G. 2021. The Computational Complexity of Understanding Binary Classifier Decisions. *J. Artif. Intell. Res.*, 70: 351–387.

Watson, D.; Gultchin, L.; Taly, A.; and Floridi, L. 2021. Local Explanations via Necessity and Sufficiency: Unifying Theory and Practice. arXiv:2103.14651.

Young, H. 1985. Monotonic Solutions of Cooperative Games. *International Journal of Game Theory*, 14: 65–72.

Yu, J.; Ignatiev, A.; and Stuckey, P. J. 2023. On Formal Feature Attribution and Its Approximation. *CoRR*, abs/2307.03380.