High Dimensional Model Explanations: an Axiomatic Approach

Neel Patel
National University of Singapore
neeltuk@gmail.com

Martin Strobel
National University of Singapore
mstrobel@comp.nus.edu.sg

Yair Zick
National University of Singapore
dcsyaz@nus.edu.sg

ABSTRACT

Complex machine learning models are regularly used in critical decision-making domains, as black-box algorithms. This has given rise to several calls for algorithmic explainability. Many explanation algorithms proposed in literature assign importance to each feature individually. However, such explanations fail to capture the joint effects of sets of features. Indeed, few works so far formally analyze *high dimensional model explanations*. In this paper, we propose a novel high dimension model explanation method that captures the joint effect of feature subsets.

We propose a new axiomatization for a generalization of the Banzhaf Index; our method can also be thought of as an approximation of a black-box model by a higher-order polynomial. In other words, this work justifies the use of the generalized Banzhaf index as a model explanation by showing that it uniquely satisfies a set of natural desiderata and that it is the optimal local approximation of a black-box model.

1 INTRODUCTION

Machine learning models are currently applied in a variety of high-stakes domains, such as healthcare, insurance, credit decisions and more. These domains require high prediction accuracy over high-dimensional data, and thus require the adoption of increasingly complex models. The ability to correctly interpret a prediction of the model's output is extremely important; however, due to the complex structure of such algorithms, they lack interpretability and transparency. The problem of ML interpretability has received a lot of attention in the machine learning community, and a wide range of explanation mechanisms have been proposed in the past few years. Broadly speaking, model explanations are based on a labeled dataset as well as other potential inputs. Most model explanation techniques focus on attribute-based explanations: for each feature i, the model explanation outputs a value ϕ_i which signifies the importance of the i-th feature in determining model predictions.

The value ϕ_i can be thought of as a score — 'Alice's high income is highly indicative of her receiving a loan' — or a counterfactual — 'had Alice's income been lower by \$10,000/year her loan would have been rejected'. Either way, the basic premise of attribute-based model explanations is to explain complex model decisions via a list of n numerical values, where n is the number of data features. Crucially, this approach fails to capture *feature interactions*. Features are often strongly intertwined, *especially* in complex ML models. For example: consider a black-box model that predicts sentiments associated with a paragraph of text. In such texts; there can be a high negative interaction effect between "not" and "bad", which attribute-based model explanations will fail to capture: assigning influence to "bad" and "not" individually can be misleading.

In this paper, we propose an axiomatic of a high dimensional model explanation method, which captures how feature interactions influence model decision-making. Our feature interaction measure explores the idea of interaction among players in a cooperative game. Power indices, for example the Shapley value [22], of cooperative games have been used extensively as feature-based model explanations [7, 9, 18]. However, the axioms characterizing power indices and interaction indices known from cooperative game theory might not be *intuitive* for explaining black-box decisions. We propose a minimal set of more natural axioms that are uniquely satisfied by an explanation method that coincides with the *Banzhaf interaction index (BII)* [12].

1.1 Our Contribution

In this work, we propose a method for high-dimensional explanations for black-box models. Our main goal is to axiomatically capture higher-order feature interaction. Our main contribution is twofold: first, we extend the idea of feature-based model explanations, which can be thought of as a local linear approximations of black-box models, to higher-order polynomial approximations. Especially, we show that our proposed measure can be obtained by approximating the black-box model by a higher-order polynomial (Section 4).

Second, we obtain a new axiomatization of the *Banzhaf interaction index* which uniquely satisfies symmetry, limit condition, general-2-efficiency, and a newly proposed axiom (in the context of Banzhaf indices): monotonicity.

Monotonicity is a rather general property which essentially means that the model explanation should change in a manner faithful to the underlying data. This is very fundamental property for an interaction measure: which states that the net contribution of the subset of features for the machine learning model f is more than that for the model g; then the interaction measure for those features for model f should be more than the interaction measure for those features for model g. Our proposed notion of monotonicity is the extension of the strong monotonicity proposed for a solution of cooperative games to higher-order coalitions. In fact, our notion coincides with strong monotonicity for $|S| = 1$.

1.2 Related Work

Model Explanations: many model explanation methods have been proposed in recent years. Some techniques provide record-based explanations [15], or generate explanations from source code [8], but the bulk of the literature on the topic focuses on feature based model explanations [2, 4, 7, 9, 24, 26]. Ancona et al. [3] offer an overview of feature based model explanations for deep neural networks. The connection to *cooperative game theory* has been

2nd Games, Agents, and Incentives Workshop (GAIW@AAMAS), 2020

widely discussed and exploited in order to generate model explanations [2, 7, 9]. With a special focus on the Shapley value and its variants [22].

Interaction Index for Cooperative Games: Two widely accepted measures of marginal influence from cooperative game theory the Shapley value [22] and the Banzhaf value [5], are uniquely derived from a set of natural axioms. Young [30] proposes monotonic solutions for cooperative games and characterizes the Shapley value — which uniquely follows strong monotonicity, symmetry, and efficiency. These measures do not capture player interactions; rather, they assign weights to individual players. Owen [19] studies pairwise interaction between players and proposes the first higher-order solution for a cooperative game. Grabisch and Roubens [12] extend it to an interaction between any subset of players and build an axiomatic foundation deriving the Shapley and the Banzhaf interaction indices. In a recent paper, Agarwal et al. [1] propose a new axiomatization for interaction among players which is inspired by the Taylor approximation of a Boolean function.

Interaction among features: interaction among features has been discussed in different communities. In statistics, there exists a vast classic literature on ANOVA based interaction among features [10, 11]. Some recent work in the deep learning literature discusses the interaction among features: Tsang et al. [28] learns interactions by inspecting the inter-layer weight matrices of a neural network, Tsang et al. [27] construct a generalized additive model that contains interaction information among features. In another line of work, Cui et al. [6], Greenside et al. [13] compute the interaction among features by computing the (expected) Hessian; this can be thought as an extension of gradient-based influence measures for neural networks [3]. Datta et al. [9] also propose an influence measure for a set of featrues called QII. It essentially measures the change of the output of a model when we randomly change a fixed set of features. Lundberg et al. [17] propose the Shapley interaction index as a high-dimensional model explanation specifically for tree-based models.

2 PRELIMINARIES

We denote sets with capital letters $A, B, \ldots$ and use lowercase letters for functions, scalars, and features. To minimize notation clutter, we try to omit braces for singletons, pairs and triplets, e.g. we write $f(i), S \cup i$ instead of $f(\{i\}), S \cup \{i\}$ and $S \cup ij, S \cup ijk$ instead of $S \cup \{i, j\}, S \cup \{i, j, k\}$.

Let $N = \{1, \ldots, n\}$ be the set of *features*. A *black-box model* is a function mapping a set of n-dimensional input vectors $\mathcal{X} \subseteq \mathbb{R}^n$ to $\mathbb{R}$. For example, a model may be given as input the details of a loan applicant (e.g. their monthly income, loan default history, etc.), and output a numerical value corresponding to the interest rate the bank should offer them on their loan application. Our objective is to generate a *model explanation* for a given *point of interest* (POI) $\vec{x} \in \mathbb{R}^n$; this explanation should (ideally) offer stakeholders some insight into the underlying decision-making process that ultimately resulted in the outcome they receive. In this work, we are interested in measuring the extent to which features, and *their high-order interactions* affect model decisions. In order to measure feature interaction effects, we adopt the *baseline comparison* approach [18, 25]; in other words, we assume the existence of a *baseline vector* $\vec{x}'$, to which we compare an input vector $\vec{x}$, in order to generate a model explanation. For example, in the automatic loan acceptance/rejection domain $\vec{x}'$ could correspond to an all-zero vector (e.g. measuring the effect of an applicant having no money in their bank account, as opposed to the true amount they have), or a vector of mean values (e.g. comparing the applicant's true income to the average population income).

In order to generate model explanations, we need to formally reason about the effect of changing features in the POI $\vec{x}$ to their baseline values. Generally speaking, changing a single feature may have no significant effect on the model prediction. For example, if $f(income = 20k, debt = 90k) = 1$, it may well be the case that changing either the applicant's low income ($20k$) or their high debt ($90k$) would not result in them receiving the loan, however, it is unreasonable to claim that neither had an effect on the outcome. To formally reason about the joint effect of features, we define a function measuring their value as a set. Given a point of interest $\vec{x}$, we define a set function as

$$v(S, \vec{x}, \vec{x}', f) = f(\vec{x}_S, \vec{x}'_{N \setminus S}) - f(\vec{x}'). \tag{1}$$

In other words, the value we assign to a set of features S is the extent to which they cause the model prediction to deviate from the baseline prediction; as a sanity check, we note that

$$v(\emptyset, \vec{x}, \vec{x}', f) = f(\vec{x}') - f(\vec{x}') = 0, \text{ and}$$
$$v(N, \vec{x}, \vec{x}', f) = f(\vec{x}) - f(\vec{x}').$$

This formulation induces a *cooperative game*, where features correspond to players. We refer to the game defined in (1) as the *feature effect game*. When clear from context, we omit $\vec{x}, \vec{x}'$ and f, focusing solely on the set of features S.

We often replace sets of features with a single feature demarcating the entire set: given a set $T \subseteq N$, $[T]$ denotes a *single* feature corresponding to the set. The *reduced game* w.r.t. the nonempty $T \subseteq N$ is defined on the features $N \setminus T \cup [T]$ with the characteristic function $v_{[T]}(\cdot, f) : 2^{N \setminus T \cup \{[T]\}} \to \mathbb{R}$:

$$v_{[T]}(S) = \begin{cases} v(S \cup T) & [T] \in S \\ v(S) & \text{otherwise} \end{cases}$$

Finally, given a subset $S \subseteq N$, S^n denotes the n-dimensional vector, that is 1 for $i \in S$ and zero otherwise.

Our objective is to generate *high dimensional model explanations*, i.e. functions that assign a value to every subset of features $S \subseteq N$. To do so, we define a *feature interaction index* $S \subseteq N$ for v as $I^v(S)$. In other words, under the game v — namely the game defined in (1) — should roughly capture the overall effect that the set of features S has on the value of v. Going back to the feature effect game defined in (1), $I^v(S)$ should measure the degree to which switching the value of features in S back to their baseline values affects the prediction for $\vec{x}$.

A key idea in our analysis is *marginal effect*: consider a single feature $i \in N$. Its marginal effect on a set $T \subseteq N \setminus \{i\}$ equals $v(T \cup i) - v(T)$, i.e. how much did the value of $v(T)$ change as a result of i joining the coalition T. The marginal effect of i on T is denoted $m_i(T)$. In particular, recalling that $v(S) = f(\vec{x}_S, \vec{x}'_{N \setminus S}) - f(\vec{x}')$, we have

$$m_i(T, v) = v(T \cup i) - v(T) = f(\vec{x}_{T \cup i}, \vec{x}'_{N \setminus \{T \cup i\}}) - f(\vec{x}_T, \vec{x}'_{N \setminus T}).$$

Thus, $m_i(T, v)$ is the marginal effect of knowing the feature i, given that the values of features in the set T are known. This is similar to other definitions considered in the literature [9, 18]. When considering a pair of features $i, j \in N$, how would one define their marginal effect on a coalition $T \subseteq N$? One very natural definition is to offset the marginal effect of adding both features to T by the marginal effects of adding i and j separately, i.e.

$$\begin{aligned} m_{ij}(T, v) =& (v(T \cup ij) - v(T)) \\ &- (v(T \cup i) - v(T)) \\ &- (v(T \cup j) - v(T)) \\ =& v(T \cup ij) - v(T \cup i) - v(T \cup j) + v(T) \end{aligned}$$

We can define the marginal contribution of a general $S \subseteq N$, in a similar manner. Let $m_S(T, v)$ be:

$$m_S(T, v) = \sum_{L \subseteq S} (-1)^{|S|-|L|} v(T \cup L).$$

This is also known as the S *discrete derivative of* v at T. We note that when $T = \emptyset$, $m_S(\emptyset, v)$ is the *Harsanyi dividend* of S, a well-known measure of the synergy (or surplus) generated by S [20]. For $T \neq \emptyset$, $m_S(T, v)$ can be thought of as the added value of having the coalition S form, given that the players in T have already committed to joining.

More generally, $m_S(T, v)$ represents the marginal interaction between features in S within the set of features $T \cup S$. Given a set of features $R \subseteq N$, we define the *primitive game* p^R as:

$$p^R(S) = \begin{cases} 1, & \text{if } R \subseteq S \\ 0, & \text{else} \end{cases}$$

The set $\mathcal{B}^N$ of all Boolean functions forms a vector space, and the set of primitive games $\mathcal{P}^N = \{p^R : R \subseteq N\}$ of all primitive games is an orthonormal basis of the vector space $\mathcal{B}^N$. Therefore any Boolean function v (In this context a cooperative game v) is uniquely represented by a linear combination of primitive games. In other words, given a cooperative game $v : 2^N \to \mathbb{R}$, we can uniquely write as a linear combination of primitive games:

$$v(\cdot) = \sum_{R \subseteq N} C_R p^R(\cdot) \tag{2}$$

The unique decomposition of cooperative games will prove useful in our characterization of high-dimensional model explanations, presented in Section 3.

3 CHARACTERIZING GOOD MODEL EXPLANATIONS

As we previously discussed, our objective is to identify *high-quality* model explanations. When pursuing quality metrics for a model explanation, one can take one of two approaches: either show that the model explanation is the optimal solution to some target (e.g. minimizes some loss function) [21, 24], or that it satisfies a set of desirable properties [7, 9, 24]. We take the latter approach in this work, describing a unique form for the high-dimensional explanation. In what follows, we explore the *Banzhaf interaction index* (BII) [12] as a potential method of generating high-dimensional model explanations. Given a cooperative game v, the Banzhaf Interaction Index for a subset $S \subseteq N$ is

$$\begin{aligned} I^v_{BAN}(S) =& \frac{1}{2^{n-|S|}} \sum_{T \subseteq N \setminus S} \sum_{L \subseteq S} (-1)^{|S|-|L|} v(L \cup T) \\ =& \frac{1}{2^{n-|S|}} \sum_{T \subseteq N \setminus S} m_S(T, v) \end{aligned} \tag{3}$$

In other words, $I^v_{BAN}(S)$ equals S's expected marginal contribution to a set $T \subseteq N \setminus S$, sampled uniformly at random. As Grabisch and Roubens [12] show, BII uniquely satisfies *Linearity*, *Symmetry*, *Dummy*, the *Recursive Property*, *Generalized 2-Efficiency*, and the *Limit Condition*. In Section 3.1, we propose a 'leaner' axiomatization of BII, which, as we argue, is more sensible in the model explanation setting. We show that BII is the unique measure which satisfies four natural axioms: *Symmetry*, *Generalized 2-Efficiency*, the *Limit Condition*, and *Monotonicity*. The first three axioms are fairly standard assumptions in identifying 'good' solutions, generalized to interaction indices by Grabisch and Roubens [12].

Symmetry (S): for any permutation π over N we have that $I^v(S) = I^{\pi v}(\pi S)$. Here, πS equals $\{\pi(i) : i \in S\}$, and πv is the game where the value of a coalition T equals $v(\pi^{-1}T)$. Symmetry is a natural property for any interaction measure: intuitively, it simply stipulates that features' interaction value is independent of their identity, and depends only on their intrinsic coalitional worth.

Generalized 2-Efficiency (GE): for any $i, j \in N$, and for any $S \subseteq N \setminus ij$:

$$I^{v[ij]}(S \cup [ij]) = I^v(S \cup i) + I^v(S \cup j)$$

Intuitively, merging two features into one feature encoding the same information results in no additional influence. Generalized 2-Efficiency extends the 2-Efficiency axiom proposed by Lehrer [16] to characterize the Banzhaf value.

Limit Condition(L): if N is the set of players of the game v then $I^v(N) = m_N(\emptyset, v) = \sum_{L \subseteq N} (-1)^{n-|L|} v(L)$. In other words, the interaction value of the set N equals exactly the added value of it forming, given that no subsets of players have pre-committed themselves to joining.

Now, we introduce the notion of monotonicity for interaction indices: given cooperative games v_1 and v_2 and a set of features $S \subseteq N$; the net interaction contribution of features S with the set T is captured by $m_S(T, v_i)$, for $i = 1, 2$. If $m_S(T, v_1) \geq m_S(T, v_2)$ for all $T \subseteq N \setminus S$ then the interaction value assigned to S should reflect this. This idea extends the strong monotonicity axiom proposed by Young [30], which states that if $m_i(T, v_1) \geq m_i(T, v_2)$ for all $T \subseteq N \setminus i$ then $I^{v_1}(i) \geq I^{v_2}(i)$.

Monotonicity(M): If $\forall T \subseteq N \setminus S$, $m_S(T, v_1) \geq m_S(T, v_2)$ and for some $T \subseteq N \setminus S$ strict inequality holds then $I^{v_1}(S) > I^{v_2}(S)$. Moreover, if $\forall T \subseteq N \setminus S$, $m_S(T, v_1) = m_S(T, v_2)$ then $I^{v_1}(S) = I^{v_2}(S)$.

Datta et al. [9] argue that the monotonicity axiom is better suited for charactering model explanations than the more 'standard' linearity axiom used in the classic characterization of the Shapley value [22], as well as the original BII characterization by Grabisch and Roubens [12].

In Section 3.2 we argue in more detail why these are suitable axioms for model explanations.

3.1 Characterizing Monotone High-Dimensional Model Explanations

The main result of this section is Theorem 3.1.

Theorem 3.1. *The only high-dimensional model explanation that satisfies (S),(GE),(L) and (M) is the Banzhaf Interaction Index.*

$$I^v(S) = \frac{1}{2^{n-s}} \sum_{T \subseteq N \setminus S} \sum_{L \subseteq S} (-1)^{|S|-|L|} v(L \cup T), \forall S \subseteq N$$

Before we prove Theorem 3.1, we require some technical claims which will be useful for our characterization result.

Proposition 3.1. *Given a primitive game p^R, for any $S \subseteq N$: if $S \nsubseteq R$ then $\forall T \subseteq N \setminus S$, $m_S(T, p^R) = 0$. In particular, $I^{p^R}_{BAN}(S) = 0$.*

Proof. Suppose that $S \nsubseteq R$. We distinguish between two cases.

Case 1: $R \nsubseteq T \cup S$. In this case, $\forall L \subseteq S$, $T \cup L$ does not contain R, thus $p^R(T \cup L) = 0$ which implies $m_S(T, p^R) = 0$.
Case 2: $R \subseteq T \cup S$. In this case, $m_S(T, p^R)$ equals

$$\sum_{L \subseteq S} (-1)^{|S|-|L|} p^R(L \cup T) = \sum_{L \subseteq S : S \cap R \subseteq L} (-1)^{|S|-|L|} =$$
$$\sum_{L \subseteq S \setminus R} (-1)^{|S|-|S \cap R|-|L|} = \sum_{L \subseteq S \setminus R} (-1)^{|S \setminus R|-|L|} =$$
$$\sum_{k=0}^{|S \setminus R|} (-1)^k \binom{|S \setminus R|}{k} = 0$$

Thus in either case $m_S(T, p^R) = 0$, and we are done. □

We note that Proposition 3.1 immediately holds for any game that is a scalar multiple of a primitive game, i.e. for any $v = c \times p^R$, and any $S \nsubseteq R$, $I^v_{BAN}(S) = 0$.

Proposition 3.2. *If $v = c \times p^R$, and $S \subseteq R$ then*

$$I^v_{BAN}(S) = \frac{c}{2^{|R|-|S|}}.$$

Proof. Since $S \subseteq R$, then for any $L \subset$ and any $T \subseteq N \setminus S$, $v(L \cup T) = 0$. Therefore,

$$I^v_{BAN}(S) = \frac{1}{2^{n-|S|}} \sum_{T \subseteq N \setminus S} \sum_{L \subseteq S} (-1)^{|S|-|L|} v(L \cup T)$$
$$= \frac{1}{2^{n-|S|}} \sum_{T \subseteq N \setminus S} v(S \cup T) \quad (4)$$

Next, if $T \subseteq N \setminus S$ does not contain $R \setminus S$, then $v(S \cup T) = 0$. Therefore, the summand in (4) equals

$$\sum_{T \subseteq N \setminus R} v(R \cup T) = c \times \sum_{T \subseteq N \setminus R} 1 = c \times 2^{n-|R|}$$

Plugging this back into (4), we obtain the desired result. □

Next, we show that the four axioms we propose imply a generalized version of the dummy property.

Lemma 3.2. *If I^v satisfies (S), (GE), (L) and (M) then if $m_S(T, v) = 0$ for all $T \subseteq N \setminus S$ then $I^v(S) = 0$.*

Proof. We begin by showing a weaker claim: if g is a *null game* where $g(S) = 0$ for all $S \subseteq N$, then $I^g(S) = 0$ for all $S \subseteq N$. For a given null game g, we have $m_S(T, g) = 0$ for all $S \subseteq N$ and $T \subseteq N \setminus S$. By symmetry, $I^g(S_1) = I^g(S_2)$ for all $S_1, S_2 \subseteq N$ with $|S_1| = |S_2|$, because for any permutation π, $g = \pi g$. Also, by the Limit Condition(L), $I^g(N) = m_N(\emptyset, g) = 0$. Similarly, for all $i_1 \neq i_2 \in N$, $I^{g[i_1, i_2]}(N \setminus i_1 i_2 \cup [i_1, i_2]) = 0$. In fact, this property holds for all $V \subset N$: $I^{g[V]}(N \setminus V \cup [V]) = 0$. Now we use (GE) property for $I^{g[V]}(N \setminus V \cup [V])$ by sequentially removing all $k \in V$ until it becomes a singleton. First, for all $k_1 \in V$;

$$0 = I^{g[V]}(N \setminus V \cup [V]) = I^{g[V \setminus k_1]}(N \setminus V \cup [V \setminus k_1]) +$$
$$I^{g[V \setminus k_1]}(N \setminus V \cup k_1) = 2 I^{g[V \setminus k_1]}(N \setminus V \cup [V \setminus k_1])$$

The second equality holds because of symmetry (S) property for the game $g_{[V \setminus k_1]}$. Now for $k_2 \in V \setminus k_1$; we similarly use the (GE) property to obtain $I^{g[V \setminus k_1]}(N \setminus V \cup [V \setminus k_1]) = 2 I^{g[V \setminus k_1]}(N \setminus V \cup [V \setminus k_1 k_2])$. We repeat this argument until only one element is left. We get $0 = I^{g[V]}(N \setminus V \cup [V]) = 2^{|V|-1} I^g(N \setminus V \cup k)$. This equality holds for all $V \subseteq N$ and all $k \in V$. Which shows that for all $S \subseteq N$, $I^g(S) = 0$.
Now, to prove the second part of the lemma, consider any game v. If $m_S(T, v) = 0$ for all $T \subseteq N \setminus S$, then $m_S(T, v) = m_S(T, g)$ for all $T \subseteq N \setminus S$, therefore by the Monotonicity property(M), $I^v(S) = I^g(S) = 0$, which concludes the proof. □

Next, let us characterize how influence measures satisfying our axioms behave on primitive games. Note that Lemma 3.3 offers a special case of Proposition 3.1 for *any* influence measure, rather than just for BII.

Lemma 3.3. *If I^v satisfies (S), (GE), (L) and (M) then for $v = c \times p^R$, $I^v(R) = c$*

Proof. We prove this lemma by inductively removing a feature $k \in N \setminus R$ and using the (GE) property at each step. Take any feature $i \in R$ and remove it from R and define $S := R \setminus \{i\}$. Now for any feature $j_1 \in N \setminus R \neq i$, by (GE) property we can write,

$$I^{v_{[ij_1]}}(S \cup [ij_1]) = I^v(S \cup i) + I^v(S \cup j_1)$$

Since $S \cup j_1 \nsubseteq R$, by Proposition 3.1, $m_{S \cup j_1}(T, v) = 0$ for all $T \subseteq N \setminus \{S \cup j_1\}$. Therefore by Lemma 3.2, $I^v(S \cup j_1) = 0$, which yields $I^v(R) = I^{v_{[ij_1]}}(S \cup [ij_1])$. We next remove $j_2 \neq j_1 \in N \setminus R$, and again invoke the (GE) property:

$$I^{v_{[ij_1 j_2]}}(S \cup [ij_1 j_2]) = I^{v_{[ij_1]}}(S \cup [ij_1]) + I^{v_{[ij_1]}}(S \cup j_2)$$

It is easy to check that $I^{v_{[ij_1]}}(S \cup j_2) = 0$ by Proposition 3.1 and Lemma 3.2 for the reduced game $v_{[ij_1]}$.

$$v_{[ij_1]}(S') = \begin{cases} c, & \text{if } S \cup [ij_1] \subseteq S' \\ 0, & \text{else} \end{cases}$$

$m_{S \cup j_2}(T, v_{[ij_1]}) = 0$ for any $T \subseteq \{N \setminus \{i, j_1\} \cup [ij_1]\} \setminus \{S \cup j_2\}$. This implies that $I^v(R) = I^{v_{[ij_1 j_2]}}(S \cup [ij_1 j_2])$. By repeating this argument for all $j \in N \setminus R$, we will have $I^v(R) = I^{v_{[N \setminus S]}}(S \cup [N \setminus S])$. We can write the reduced game $v_{[N \setminus S]}$ as

$$v_{[N \setminus S]}(S') = \begin{cases} c, & \text{if } S' = S \cup [N \setminus S] \\ 0, & \text{else.} \end{cases}$$

By the Limit (L) property,

$$I^{v_{[N\setminus S]}}(S \cup [N \setminus S]) = m_{S\cup[N\setminus S]}(\emptyset, v_{[N\setminus S]}) = c,$$

which concludes the proof □

We are now ready to characterize interaction indices that uniquely satisfy the above properties.

Proof of Theorem 3.1. We first show that BII satisfies all the properties. BII trivially satisfies (S), (L) and (M). To show that it satisfies (GE), take any $S \subseteq N \setminus ij$

$$\begin{aligned} I^v_{BAN}(S \cup i) =& \frac{1}{2^{n-s-1}} \sum_{T\subseteq N\setminus(S\cup i)} m_{S\cup i}(T, v) \\ =& \frac{1}{2^{n-s-1}} \sum_{T\subseteq N\setminus(S\cup ij)} [m_S(T \cup i, v) - m_S(T, v)] \\ &+ \frac{1}{2^{n-s-1}} \sum_{T\subseteq N\setminus(S\cup ij)} [m_S(T \cup ij, v) - m_S(T \cup j, v)] \end{aligned}$$

A similar calculation for j shows that

$$\begin{aligned} I^v_{BAN}(S \cup j) =& \frac{1}{2^{n-s-1}} \sum_{T\subseteq N\setminus(S\cup ij)} [m_S(T \cup j, v) - m_S(T, v)] \\ &+ \frac{1}{2^{n-s-1}} \sum_{T\subseteq N\setminus(S\cup ij)} [m_S(T \cup ij, v) - m_S(T \cup i, v)] \end{aligned}$$

Thus, $I^v_{BAN}(S \cup i) + I^v_{BAN}(S \cup j)$ equals

$$\frac{1}{2^{n-s-2}} \times \sum_{T\subseteq N\setminus(S\cup ij)} [m_S(T \cup ij, f) - m_S(T, f)] \tag{5}$$

Equation (5) shows that $I^{v_{[ij]}}(S \cup [ij]) = I^v(S \cup i) + I^v(S \cup j)$.
BII satisfies the four axioms; to show that it *uniquely* satisfies them, we use the fact that v can be uniquely expressed as the sum of primitive games;

$$v = \sum_{R\subseteq N} C_R p^R \tag{6}$$

We define the *index* Γ of a cooperative game v to be the minimum number of terms in the expression of the form (6). We prove the theorem by induction on Γ. For $\Gamma = 0$, in Lemma 3.2, $I^v(S) = 0$ for all $S \subseteq N$, which coincides with the Banzhaf interaction index.
If $\Gamma = 1$ then $v = C_R p^R$ for some $R \subseteq N$. Consider $S \nsubseteq R$; Proposition 3.1 implies that $m_S(T, v) = 0$ for all $T \subseteq N$, which in turn implies $I^v(S) = 0 = I^v_{BAN}(S)$. By Lemma 3.3, $I^v(R) = C_R$, which equals $I^v_{BAN}(R)$ by Proposition 3.2. To complete the proof for the first inductive step, we need to show that for all $S \subseteq R$, $I^v(S) = I^v_{BAN}(S) = \frac{C_R}{2^{|R|-|S|}}$. If $S_1, S_2 \subseteq R$ and $s_1 = s_2$ then by symmetry, $I^v(S_1) = I^v(S_2)$: we can define a permutation π over N such that S_2 bijectively maps to some S_1, and all $i \notin S_1 \cup S_2$: are invariant. By the Symmetry property $I^v(S_2) = I^{\pi v}(S_1)$; however, $\pi v = v$ because $v = C_R p^R$. Now, consider any $i_1 \neq i_2 \in R$ and define $S := R \setminus \{i_1, i_2\}$; by the GE property, we can write

$$I^{v_{[i_1 i_2]}}(S \cup [i_1 i_2]) = I^v(S \cup i_1) + I^v(S \cup i_2)$$

which implies for any $Q \subset R$ with $|Q| = |R|-1$, $I^v(Q) = \frac{1}{2} I^{v_{[i_1 i_2]}}(S \cup [i_1 i_2])$. The reduced game $v_{[i_1 i_2]}$ is

$$v_{[i_1 i_2]}(S') = \begin{cases} C_R, & \text{if } S \cup [i_1 i_2] \subseteq S' \\ 0, & \text{else} \end{cases}$$

By Lemma 3.3, $I^{v_{[i_1 i_2]}}(S \cup [i_1 i_2]) = C_R$ and $I^v(Q) = \frac{C_R}{2}$. This property holds for all $T \subset N$, $v_{[T]}$; $I^{v_{[T]}}(N \setminus T \cup [T]) = C_R$. By inductively using the (GE) property, in a manner similar to Lemma 3.3, we show that $I^v(Q) = \frac{C_R}{2^{|R|-|Q|}}$. By Proposition 3.2, this coincides with Banzhaf interaction index concluding the first inductive step. To complete the proof, assume that $I^v(S)$ coincides with the Banzhaf interaction index whenever the index of the game v is at most $\Gamma = \gamma$. Suppose that v has an index $\gamma + 1$, and expressed as

$$v = \sum_{k=1}^{\gamma+1} C_{R_k} p^{R_k}$$

Let $R = \bigcap_{k=0}^{\gamma+1} R_k$, and suppose that $S \nsubseteq R$. We define another game w:

$$w = \sum_{k:S\subseteq R_k} C_{R_k} p^{R_k}$$

Since $S \nsubseteq R$, the index of w is strictly smaller than $\gamma + 1$. We claim that for all $T \subseteq N \setminus S$; $m_S(T, v) = m_S(T, w)$. Indeed, consider any $T \subseteq N \setminus S$; $m_S(T, v)$ equals

$$\begin{aligned} \sum_{L\subseteq S} (-1)^{|S|-|L|} v(T \cup L) = \sum_{L\subseteq S} (-1)^{|S|-|L|} \sum_{k=1}^{\gamma+1} C_{R_k} p^{R_k}(T \cup L) = \\ \sum_{k=1}^{\gamma+1} \sum_{L\subseteq S} (-1)^{|S|-|L|} C_{R_k} p^{R_k}(T \cup L) = \sum_{k=1}^{\gamma+1} C_{R_k} m_S(T, p^{R_k}) = \\ \sum_{k:S\subseteq R_k} C_{R_k} m_S(T, p^{R_k}) = m_S(T, w) \end{aligned}$$

The second-last equality holds by Proposition 3.1, hence by induction on Γ and monotonicity(M) $I^v(S)$ coincides with BII for all $S \nsubseteq R$.
It remains to show that $I^v(S)$ coincides with BII when $S \subseteq R$. for any $S \subseteq R$, consider any $i \in S$ and define $S' := S \setminus i$. Take any $j \in N$ such that $j_1 \in R_1 \setminus R_2 \cup R_2 \setminus R_1$. By the (GE) property, we can write

$$I^v(S) = I^{v_{[ij_1]}}(S' \cup [ij_1]) - I^v(S' \cup j_1) \tag{7}$$

In Equation (7), $S' \cup j_1 \nsubseteq R$, therefore as previously shown, $I^v(S' \cup j_1)$ coincides with BII for the game v. Consider the restricted game $v_{[ij_1]}$:

$$v_{[ij]} = \sum_{k=0}^{\gamma+1} C_{R_k} p^{R_k \setminus ij_1 \cup [ij_1]}_{[ij_1]}$$

Consider $j_2 \neq j_1 \in R_1 \setminus R_2 \cup R_2 \setminus R_1$. By the (GE) property,

$$I^{v_{[ij_1 j_2]}}(S' \cup [ij_1 j_2]) = I^{v_{[ij_1]}}(S' \cup [ij_1]) + I^{v_{[ij_1]}}(S' \cup j_2) \tag{8}$$

In Equation (8), $S' \cup j_2 \nsubseteq \bigcap_{k=1}^{I+1} R_k \setminus ij_1 \cup [ij_1]$, therefore as we have shown before, $I^{v_{[ij_1]}}(S' \cup j_2) = I^{v_{[ij_1]}}_{BAN}(S' \cup j_2)$. Let us denote $T = R_1 \setminus R_2 \cup R_2 \setminus R_1$ and $T' = i \cup T$. By repeating this argument for all $j_3, \ldots, j_t \in T$ and exploiting the (GE) property for each j_ℓ, we can write $I^v(S)$ as:

$$I^v(S) = I^{v_{[T']}}(S' \cup [T']) - I^v(S' \cup j_1) - \sum_{l=1}^{t} I^{v_{[ij_1 \ldots j_\ell]}}(S' \cup j_\ell) \tag{9}$$

All of the summands in (9) coincide with BII, because $S' \cup j_\ell \notin \bigcap_{k=1}^{I+1} R_k \setminus ij_1 \dots j_{\ell-1} \cup [ij_1 \dots j_{\ell-1}]$ for all $\ell = 1, \dots, t$. We can write the reduced game $v_{[T']}$ as

$$v_{[T']} = (C_{R_1} + C_{R_2}) p_{[T']}^{(R_1 \cap R_2 \setminus i) \cup [T']} + \sum_{k=3}^{\gamma+1} C_{R_k} p_{[T']}^{(R_k \setminus T') \cup [T']}$$

Thus, the index of the reduced game $v_{[T']}$ is strictly smaller than $\gamma+1$. By induction, $I^{v_{[T']}}(S' \cup [V])$ coincides with BII for the reduced game $v_{[T']}$. $I^v(S)$ can be written as

$$I^v(S) = I^{v_{[T']}}(S' \cup [T']) - \sum_{l=1}^{t} I^{v_{[ij_1 \dots j_\ell]}}(S' \cup j_\ell)$$

By using the (GE) property inductively, $I^v_{BAN}(S)$ can also be written in the same form, which implies that $I^v(S)$ coincides with the Banzhaf interaction index for all, $S \subseteq R$. □

3.2 Explaining Our Model Explanations

Does the BII measure make sense in the model explanation domain? This is purely a function of the strength of the axioms we set forth. Symmetry is natural enough: if a model explanation depends on the indices of its features then it fails a basic validity test. The index in which a feature appears has no bearing on the underlying trained model (ideally), nor does it affect the outcome.

Recall that Generalized Efficiency requires that model explanations should be invariant under feature merging. In other words - artificially treating a pair of features as a single entity (while maintaining the same underlying model) should not have any effect on how feature behaviors are explained. Interestingly, Shapley values are not invariant under feature merging, a result shown by Lehrer [16]. The following examples illustrate what this entails in actual applications.

Example 3.4. Consider an sentiment analysis task where a model predict if a movie review was positive. In a preliminary step the text is parsed by a parser to be machine readable. This can be done in many different ways. For example the sentence "This isn't a absolutely terrible movie" Can be parsed as

| This | isn't | a | absolutely | terrible | movie | . |

or as

| This | is | n't | a | absolut | ely | terrible | movie | . |.

Generalized Efficiency ensures that the influences of "is" and "n't" in the second version add up to the influence of "isn't" in the first. In other words, Generalized Efficiency ensures that the influence of features generated through different parsers behaves in a sensible manner.

Example 3.5. Features might be "merged" in another situation when features that were readily available during the training of a model end up being costly to obtain during its deployment. If additionally these features are highly correlated with other features they might just be coupled. E.g. generally birds can fly, so the features IS_BIRD and CAN_FLY may simply be merged at prediction time[1], to make it easier to enter information into a classifier. Again, Generalized Efficiency ensures that the influence of the merged feature relates in a natural way to the influence of the original features.

The Limit condition normalizes the overall influence to be the discrete derivative of $v(\cdot, f)$ with respect to N. In other words, the total influence distributed to sets of features equals the total marginal effect of reverting features to their baseline values. This is an interesting departure from other efficiency measures. Shapley-based measures require efficiency with respect to $f(\vec{x})$ (or variants thereof), i.e. the total amount of influence should equal the total value the classifier takes at the point of interest (or the difference between the classifier and the baseline value). We require that the total influence equals the (discretized) rate in which features change the outcome. This makes BII more similar in spirit to gradient based model explanations, which are often used as the basic mechanism for generating model explanations in several application domains [23].

Monotonicity is a very natural property in the model explanation domain: if a set of features has a greater effect on the value $f(\vec{x})$, this should be reflected in the amount of influence one attributes to it. This has already been established in prior works, for Shapley-based measures [9, 18]. However, this property does not naturally generalize when using Shapley-based high-dimensional model explanations. Agarwal et al. [1] propose a novel generalization of the Shapley value to high-dimensional model explanations, which fails monotonicity for smaller interactions (size of $< k$) for $k-$th order explanation, however, interactions of size k follow monotonicity.

Example 3.6. Given a function $f_c(x_1, x_2, x_3) = cx_1x_2x_3$ with $c > 0$ defined on binary input space (for example, f is the result of an image classification task where x_i denotes the presence/absence of particular super-pixel). We assume that the baseline is $\vec{x}' = (0, 0, 0)$. Thus, $v(S, \vec{x}, \vec{x}', f_c) = 0$ if $\{2, 3\} \nsubseteq S$, resulting in $v(\{2, 3\}, f_c) = 0$ and $v(N, f_c) = v(\{1, 2, 3\}, f_c) = cx_1x_2x_3$. What is the interaction value between 1 and 2? Intuitively $\{1, 2\}$ offer some degree of interaction that monotonically grows as c increases. Moreover, it is easy to see that $v(\cdot, f_c) \geq v(\cdot, f_{c'})$ whenever $c \geq c'$. Set-QII fails to satisfies the monotoncity, and fails to capture the interaction between $\{1, 2\}$ for any c. Set-QII$(\{1, 2\}, S, f_c) = 0$ for all c.

Similarly, the Shapley-Taylor interaction index for $S = \{1, 2\}$ and $k = 3$ is 0 as it does not follow the monotonicity property, however for $k = 2$ it satisfies monotonicity and interaction value for $\{1, 2\}$ is $\frac{c}{3}$. The BII value for $\{1, 2\}$ is c.

In the next example, we demonstrate that the Shapley interaction index can be misleading in simple situations. We consider the general majority classification function which exhibits pairwise feature interaction. Shapley interaction indices fail to capture these interactions. Moreover, Shapley-Taylor interaction indices fail to capture the sign of pairwise interactions for the same function.

Example 3.7. Consider a classification function whose input space is binary. Let the classification function f be: $f(x_1, \dots, x_n) = 1$ iff $\sum_i x_i \geq k$ and 0 elsewhere with the baseline vector $\vec{x}' = \vec{0}$. Thus, $v(S, \vec{1}, \vec{x}') = 1$ iff $|S| \geq k$ and 0 otherwise. For $k = \frac{n}{2}$, the function coincides with the majority function discussed in the cooperative game theory literature. Clearly, there exists pairwise interaction among features, however, the Shapley interaction value for each pairwise feature is 0. In contrast, the pairwise Banzhaf

[1]The authors are aware of the existence of ostriches, emus, penguins and the fearsome cassowary.

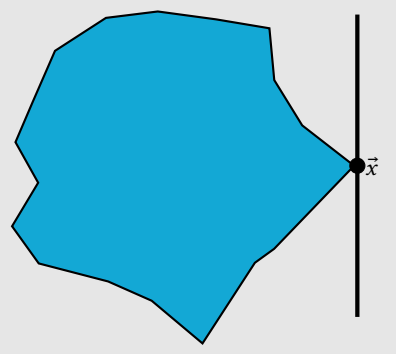

Figure 1: A scenario where a linear approximation cannot explain the complex model's behavior well.

interaction index for any feature pair $\{i, j\}$ is $c_k(2k-(n+3))$ where $c_k > 0$. Pairwise interaction is negative when $2k < n+3$. This can be explained by the following argument: the number of winning coalitions containing $\{i, j\}$ is $\binom{n-2}{k-2}$, and the number of winning coalitions that do not contain $\{i, j\}$ $\binom{n-2}{k}$, which is higher for smaller k. This shows that $\{i, j\}$ has more interaction effect for the output 0. On the other hand, the Shapley-Taylor pairwise interaction index is $\frac{2}{n(n-1)}$, and fails to capture the sign of the interaction index.

In this section, we argued about our proposed model explanation *makes sense* for capturing feature interaction in black-box decision making. In Section 4, we show that BII can be interpreted as a polynomial approximation, offering additional intuition as to why our explanation method is *good*.

4 GEOMETRICAL INTERACTION AND BII

The geometry of model explanation is relatively well understood for attribute-based methods [18, 21, 24]; Ribeiro et al. [21] suggest that attribute-based explanation methods can be thought of a local linear approximation of a black-box function $f(\cdot)$ around a point of interest (POI) $\vec{x}$. Linear attribution methods take the following form:

$$g(\vec{x}) = I_0 + \sum_{i=1}^{N} I_i x_i \tag{10}$$

In Equation (10), I_i captures the importance of feature i. The major problem with these methods is that the underlying black-box model can be extremely non-linear around POI $\vec{x}$. In those cases, the explanation fails to approximate the black-box model f. Figure 1 illustrates such a scenario. We note that the point of interest in Figure 1 has no particular linear local model that well approximates the true model; this is not unusual when considering model outliers. What's worse, outliers are often the points that need to be explained the most. In order to better capture the behavior of a black-box model f, we can naturally consider a higher-order polynomial as a local approximation instated of a simple local linear approximation. For better visualization, we first assume that the black-box model $f : \{0, 1\}^N \to \mathbb{R}$ takes a binary input vector mainly referred to as the humanly understandable feature representation [18, 21]. Interaction among a set of features can be thought of as a higher-order polynomial approximation extending the attribute-based explanation. First, we start with quadratic approximation of the black-box model $f(\cdot)$,

$$g^k(\vec{x}) = I_0 + \sum_{i=1}^{N} I(i)x_i + \sum_{i<j} I(\{i, j\})x_i x_j \tag{11}$$

In Equation (11), $I(\{i, j\})$ captures the interaction between feature i and j; $I(i), I(j)$ capture the importance of i and j, as they do in Equation (10). Thus, it is not unreasonable to assume that $I(\{i, j\})$ capture the pure interaction effect of i and j: we can delegate the singular effects to $I(i)$, having the resultant coefficient of $x_i x_j$ capture the 'pure' interaction between i and j. For instance, consider a sentiment analysis problem, both the tokens "bad" and "not" have negative influence on the machine learning task. However, when they are present together as "not bad", their influence is positive. In this simple example, it would be desirable to have I("not") and I("bad") < 0, but $I(\{$"not","bad"$\}) > 0$. The idea of higher order interactions can be extended similarly.

Consider a global polynomial approximation of $f(\cdot)$ by a k-degree polynomial in Equation (12)

$$g^k(\vec{x}) = I_0 + \sum_{S\subset N;|S'|<k} \left(I(S')\prod_{j\in S'} x_j\right) + \sum_{S\subset N;|S|=k} \left(I(S)\prod_{j\in S} x_j\right) \tag{12}$$

Again, to capture interaction among the set of features S such that $|S| = k$, we should remove all internal interaction effects captured by $I(S')$ for $S' \subset S$.

Therefore in Equation 12; $I(S)$ for $|S| = k$ can be thought of an interaction effect of subset of features S for the underlying black-box model $f(\cdot)$. The polynomial g^k is meant to locally approximate f around the POI $\vec{x}$; what is the best approximation? Finding the best fitting polynomial of the highest possible degree seems like a natural objective. However, we argue that taking this approach runs the risk of ignoring lower order feature interactions and their possible effects.

Example 4.1. Consider the degree 3 polynomial studied in Example 3.6, $f(x_1, x_2, x_3) = cx_1x_2x_3$ with the baseline set to $(1, 1, 1)$ (rather than $(0, 0, 0)$ as was the case in Example 3.6). The best approximation to f is clearly itself. However, if we do so, then the interaction coefficients for variable pairs will be zero. This is arguably undesirable: for example, if $\vec{x} = (0, 0, 1)$, then x_3 has virtually no impact (it is already set at the baseline). Similarly, x_1 and x_2 have little individual effect, but do have significant joint effect - it is only when both are set to 1 that we observe any change in label.

Now we formally define the optimization problem to find the "best" k-degree polynomial approximation of the black-box model $f(\cdot)$ globally. Let $\mathcal{P}^k$ be the set of k-degree polynomials of the form

$$g^k(\vec{x}) = I_0 + \sum_{S\subset N;|S|\leq k} \left(I(S)\prod_{j\in S} x_j\right).$$

We are interested in finding a polynomial $g_f^k(\cdot)$ which globally minimizes the quadratic loss between $f(\vec{x})$ and $g(\vec{x}) \in \mathcal{P}^k$ for all $\vec{x} \in 2^N$, i.e.

$$g_f^k(\vec{x}) = \arg\min_{g(\cdot)\in\mathcal{P}^k} \sum_{\vec{x}\in 2^N} [f(\vec{x}) - g(\vec{x})]^2 \tag{13}$$

The interaction among features in S with $|S| = k$ is measured as the coefficient of $\prod_{i \in S} x_i$ in the least square approximation of $f(\cdot)$ with a polynomial of degree k. Theorem 4.2 shows that this geometrical definition of feature interaction coincides with the Banzhaf interaction index.

Theorem 4.2. *Let $g_f^k(\vec{x})$ be the k-degree solution of the optimization problem in Equation* (13). *Then the coefficients of $\prod_{i \in S} x_i$ for $|S| = k$ is given by the Banzhaf interaction index (see Equation 3).*

Proof. The proof of the Theorem is a simple corollary of Hammer and Holzman [14, Theorem 4.2]. □

Theorem 4.2 provides an intuitive argument for BII being a "good" measure for capturing feature interaction.

5 CONCLUSIONS AND FUTURE WORK

We discuss the problem of identifying and measuring interactions among features in decision-making algorithms. We present a novel characterization of the Banzhaf interaction measure which uniquely satisfies a set of natural properties. In addition, it optimizes a natural objective function, providing a geometrical interpretation of our interaction measure.

Designing *provably sound* higher-order explanations for machine learning models in high stake domains is important. Axiomatizing model explanations mathematically justifies the chosen interaction measure, which helps foster trust in the explanation method. In this paper we have only demonstrated the mathematical properties of our interaction measures, however, we will add an extensive experiment section to show the effectiveness of our measure and axiomatics in the full version of this work.

We believe that the game theory/fair division community should be an active part of the discussion of algorithmic transparency. There seems to be a general receptiveness of applying cooperative solution concepts in model explanation, most prominently the Shapley value. However, other solution concepts are considered (e.g. variants of the Banzhaf index as seen in this work and others [7, 24]) and even the core [29]. More importantly, axiomatic approaches, commonly used in the economic literature, are finding their way into the field, as the need for provable trustworthiness in high-stakes ML applications grows.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number AISG-RP-2018-009).

REFERENCES

[1] Ashish Agarwal, Kedar Dhamdhere, and Mukund Sundararajan. 2019. A New Interaction Index inspired by the Taylor Series. (2019). arXiv:1902.05622
[2] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2017. A unified view of gradient-based attribution methods for deep neural networks. In *NIPS 2017-Workshop on Interpreting, Explaining and Visualizing Deep Learning*. Long Beach, CA, USA, 1–11.
[3] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. 1–16.
[4] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11, Jun (2010), 1803–1831.
[5] John F Banzhaf III. 1964. Weighted voting doesn't work: A mathematical analysis. *Rutgers L. Rev.* 19 (1964), 317.
[6] Tianyu Cui, Pekka Marttinen, and Samuel Kaski. 2019. Recovering Pairwise Interactions Using Neural Networks. (2019). arXiv:1901.08361
[7] Amit Datta, Anupam Datta, Ariel D Procaccia, and Yair Zick. 2015. Influence in classification via cooperative game theory. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*. 511–517.
[8] Anupam Datta, Matthew Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. 2017. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. In *Proceedings of the 24th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 1193–1210.
[9] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Proceedings of the 37th IEEE Symposium on Security and Privacy (Oakland)*. IEEE, 598–617.
[10] Ronald Aylmer Fisher. 1992. Statistical methods for research workers. In *Breakthroughs in statistics*. Springer, 66–70.
[11] Andrew Gelman et al. 2005. Analysis of variance—why it is more important than ever. *The annals of statistics* 33, 1 (2005), 1–53.
[12] Michel Grabisch and Marc Roubens. 1999. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory* 28, 4 (1999), 547–565.
[13] Peyton Greenside, Tyler Shimko, Polly Fordyce, and Anshul Kundaje. 2018. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics* 34, 17 (2018), i629–i637.
[14] Peter L Hammer and Ron Holzman. 1992. Approximations of pseudo-Boolean functions; applications to game theory. *Zeitschrift für Operations Research* 36, 1 (1992), 3–21.
[15] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 1885–1894.
[16] Ehud Lehrer. 1988. An axiomatization of the Banzhaf value. *International Journal of Game Theory* 17, 2 (1988), 89–99.
[17] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. 2018. Consistent individualized feature attribution for tree ensembles. (2018). arXiv:1802.03888
[18] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*. 4765–4774.
[19] Guillermo Owen. 1972. Multilinear extensions of games. *Management Science* 18, 5-part-2 (1972), 64–79.
[20] B. Peleg and P. Sudhölter. 2007. *Introduction to the Theory of Cooperative Games* (second ed.). Theory and Decision Library. Series C: Game Theory, Mathematical Programming and Operations Research, Vol. 34. Springer, Berlin.
[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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.
[22] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.
[23] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. (2013). arXiv:1312.6034
[24] Jakub Sliwinski, Martin Strobel, and Yair Zick. 2019. A Characterization of Monotone Influence Measures for Data Classification. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*. 718–725.
[25] Mukund Sundararajan and Amir Najmi. 2019. The many Shapley values for model explanation. (2019). arXiv:1908.08474
[26] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 3319–3328.
[27] Michael Tsang, Dehua Cheng, and Yan Liu. 2017. Detecting statistical interactions from neural network weights. (2017). arXiv:1705.04977
[28] Michael Tsang, Hanpeng Liu, Sanjay Purushotham, Pavankumar Murali, and Yan Liu. 2018. Neural interaction transparency (NIT): disentangling learned interactions for improved interpretability. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NIPS)*. 5804–5813.
[29] Tom Yan and Ariel D. Procaccia. 2020. If You Like Shapley Then You'll Love the Core. (2020). Working Paper.
[30] H Peyton Young. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory* 14, 2 (1985), 65–72.