The Shapley Value in Machine Learning

Benedek Rozemberczki[1] , **Lauren Watson**[2] , **Péter Bayer**[3] , **Hao-Tsung Yang**[2] , **Olivér Kiss**[4] , **Sebastian Nilsson**[1] and **Rik Sarkar**[2]
[1]Research Data & Analytics, Research & Development IT, AstraZeneca
[2]The University of Edinburgh, School of Informatics
[3] Toulouse School of Economics & Institute for Advanced Study in Toulouse
[4]Central European University, Department of Economics and Business
benedek.rozemberczki@astrazeneca.com

Abstract

Over the last few years, the Shapley value, a solution concept from cooperative game theory, has found numerous applications in machine learning. In this paper, we first discuss fundamental concepts of cooperative game theory and axiomatic properties of the Shapley value. Then, we give an overview of the most important applications of the Shapley value in machine learning: feature selection, explainability, multi-agent reinforcement learning, ensemble pruning, and data valuation. We examine the most crucial limitations of the Shapley value and point out directions for future research.

1 Introduction

Measuring importance and the attribution of various gains is a central problem in many practical aspects of machine learning such as explainability [Lundberg *et al.*, 2017], feature selection [Cohen *et al.*, 2007], data valuation [Ghorbani *et al.*, 2019], ensemble pruning [Rozemberczki *et al.*, 2021] and federated learning [Wang *et al.*, 2020; Fan *et al.*, 2021]. For example, one might ask: What is the importance of a feature in the decisions of a machine learning model? How much is an individual data point worth? Which models are the most valuable in an ensemble? These questions have been addressed in different domains using specific approaches. Interestingly, there is also a general and unified approach to these questions as a solution to a *transferable utility* (TU) cooperative game. In contrast with other approaches, solution concepts of TU games are theoretically motivated with axiomatic properties. The best known solution is the *Shapley value* [Shapley, 1953] characterized by desiderata that include fairness, symmetry, and efficiency [Chalkiadakis *et al.*, 2011].

In the TU setting, a cooperative game consists of: a *player set* and a scalar-valued *characteristic function* that defines the value of *coalitions* (subsets of players). In such a game, the Shapley value offers a rigorous and intuitive way to distribute the collective value (e.g. the revenue, profit, or cost) of the team across individuals. To apply this idea to machine learning, we need to define two components: the player set and the characteristic function. In a machine learning setting *players* may be represented by a set of input features, reinforcement learning agents, data points, models in an ensemble, or data silos. The characteristic function can then describe the goodness of fit for a model, reward in reinforcement learning, financial gain on instance level predictions, or out-of-sample model performance. We provide an example about model valuation in an ensemble [Rozemberczki *et al.*, 2021] in Figure 1.

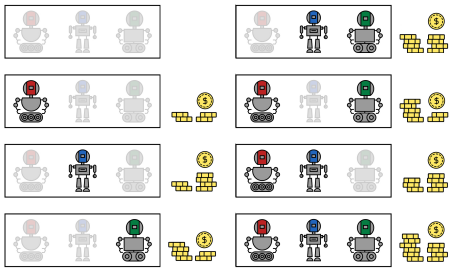

Figure 1: The Shapley value can be used to solve cooperative games. An ensemble game is a machine learning application for it – models in an ensemble are players (red, blue, and green robots) and the financial gain of the predictions is the payoff (coins) for each possible coalition (rectangles). The Shapley value can distribute the gain of the grand coalition (right bottom corner) among models.

Present work. We introduce basic definitions of cooperative games and present the Shapley value, a solution concept that can allocate gains in these games to individual players. We discuss its properties and emphasize why these are important in machine learning. We overview applications of the Shapley value in machine learning: feature selection, data valuation, explainability, reinforcement learning, and model valuation. Finally, we discuss the limitations of the Shapley value and point out future directions. The survey is supported by a collection of related work under https://github.com/AstraZeneca/awesome-shapley-value.

2 Background

This section introduces cooperative games and the Shapley value followed by its properties. We also provide an illustrative running example for our definitions.

2.1 Cooperative Games and the Shapley Value

Definition 1. ***Player set and coalitions.*** *Let $\mathcal{N} = \{1, \ldots, n\}$ be the finite set of players. We call each non-empty subset $\mathcal{S} \subseteq \mathcal{N}$ a coalition and $\mathcal{N}$ itself the grand coalition.*

Definition 2. ***Cooperative game.*** *A TU game is defined by the pair $(\mathcal{N}, v)$ where $v : 2^{\mathcal{N}} \to \mathbb{R}$ is a mapping called the*

characteristic function or the coalition function of the game assigning a real number to each coalition and satisfying $v(\emptyset) = 0$.

Example 1. *Let us consider a 3-player cooperative game where* $\mathcal{N} = \{1, 2, 3\}$. *The characteristic function defines the payoff for each coalition. Let these payoffs be given as:*

$$v(\emptyset) = 0; \quad v(\{1\}) = 7; \quad v(\{2\}) = 11; \quad v(\{3\}) = 14;$$
$$v(\{1,2\}) = 18; \quad v(\{1,3\}) = 21; \quad v(\{2,3\}) = 23; \quad v(\{1,2,3\}) = 25.$$

Definition 3. ***Set of feasible payoff vectors.*** *Let us define* $\mathcal{Z}(\mathcal{N}, v) = \{\mathbf{z} \in \mathbb{R}^{\mathcal{N}} \mid \sum_{i \in \mathcal{N}} z_i \leq v(\mathcal{N})\}$ *the set of feasible payoff vectors for the cooperative game* $(\mathcal{N}, v)$.

Definition 4. ***Solution concept and solution vector.*** *Solution concept* Φ *is a mapping associating a subset* $\Phi(\mathcal{N}, v) \subseteq \mathcal{Z}(\mathcal{N}, v)$ *to every TU game* $(\mathcal{N}, v)$. *A solution vector* $\phi(\mathcal{N}, v) \in \mathbb{R}^{\mathcal{N}}$ *to the cooperative game* $(\mathcal{N}, v)$ *satisfies solution concept* Φ *if* $\phi(\mathcal{N}, v) \in \Phi(\mathcal{N}, v)$. *Solution concept* Φ *is single-valued if for every* $(\mathcal{N}, v)$ *the set* $\Phi(\mathcal{N}, v)$ *is a singleton.*

A solution concept defines an allocation principle through which rewards can be given to the individual players. The sum of these rewards cannot exceed the value of the grand coalition $v(\mathcal{N})$. Solution vectors are specific allocations satisfying the principles of the solution concept.

Definition 5. ***Permutations of the player set.*** *Let* $\Pi(\mathcal{N})$ *be the set of all permutations defined on* $\mathcal{N}$, *a specific permutation is written as* $\pi \in \Pi(\mathcal{N})$ *and* $\pi(i)$ *is the position of player* $i \in \mathcal{N}$ *in permutation* π.

Definition 6. ***Predecessor set.*** *Let the set of predecessors of player* $i \in \mathcal{N}$ *in permutation* π *be the coalition:*

$$\mathcal{P}_i^{\pi} = \{j \in \mathcal{N} \mid \pi(j) < \pi(i)\}.$$

Let us imagine that the permutation of the players in our illustrative game is $\pi = (3, 2, 1)$. Under this permutation the predecessor set of the 1^{st} player is $\mathcal{P}_1^{\pi} = \{3, 2\}$, that of the 2^{nd} player is $\mathcal{P}_2^{\pi} = \{3\}$ and $\mathcal{P}_3^{\pi} = \emptyset$.

Definition 7. ***Shapley value.*** *The Shapley value [Shapley, 1953] is a single-valued solution concept for cooperative games. The* i^{th} *component of the single solution vector satisfying this solution concept for any cooperative game* $(\mathcal{N}, v)$ *is given by Equation 1.*

$$\phi_i^{Sh} = \frac{1}{|\Pi(\mathcal{N})|} \sum_{\pi \in \Pi(\mathcal{N})} \underbrace{[v(\mathcal{P}_i^{\pi} \cup \{i\}) - v(\mathcal{P}_i^{\pi})]}_{\substack{\textit{Player } i\textit{'s marginal contribution} \\ \textit{in permutation } \pi}} \tag{1}$$

The Shapley value of a player is the average marginal contribution of the player to the value of the predecessor set over every possible permutation of the player set. Table 1 contains manual calculations of the players' marginal contributions to each permutation and their Shapley values in Example 1.

2.2 Properties of the Shapley Value

We define the solution concept properties that characterize the Shapley value and emphasize their relevance and meaning in a *feature selection game*. In this game input features are players, coalitions are subsets of features and the payoff is a scalar valued goodness of fit for a machine learning model using these inputs.

Definition 8. ***Null player.*** *Player* i *is called a null player if* $v(\mathcal{S} \cup \{i\}) = v(\mathcal{S})\ \forall \mathcal{S} \subseteq \mathcal{N} \setminus \{i\}$. *A solution concept* Φ *satisfies the null player property if for every game* $(\mathcal{N}, v)$, *every* $\phi(\mathcal{N}, v) \in \Phi(\mathcal{N}, v)$, *and every null player* i *it holds that* $\phi_i(\mathcal{N}, v) = 0$.

	Marginal Contribution		
Permutation	**Player 1**	**Player 2**	**Player 3**
(1, 2, 3)	7	11	7
(1, 3, 2)	7	4	14
(2, 1, 3)	7	11	7
(2, 3, 1)	2	11	12
(3, 1, 2)	7	4	14
(3, 2, 1)	2	9	14
Shapley value	32/6	50/6	68/6

Table 1: The permutations of the player set, marginal contributions of the players in each permutation and the Shapley values.

In the feature selection game a solution concept with the null player property assigns zero value to those features that never increase the goodness of fit when added to the feature set.

Definition 9. ***Efficiency.*** *A solution concept* Φ *is efficient or Pareto optimal if for every game* $(\mathcal{N}, v)$ *and every solution vector* $\phi(\mathcal{N}, v) \in \Phi(\mathcal{N}, v)$ *it holds that* $\sum_{i \in \mathcal{N}} \phi_i(\mathcal{N}, v) = v(\mathcal{N})$.

Consider the goodness of fit of the model trained using the whole set of features. The importance measures assigned to features by an efficient solution concept sum to this goodness of fit. This allows for quantifying the contribution of features to the performance of a model trained on the whole feature set.

Definition 10. ***Symmetry.*** *Two players* i *and* j *are symmetric if* $v(\mathcal{S} \cup \{i\}) = v(\mathcal{S} \cup \{j\})\ \forall \mathcal{S} \subseteq \mathcal{N} \setminus \{i, j\}$. *A solution concept* Φ *satisfies symmetry if for all* $(\mathcal{N}, v)$ *for all* $\phi(\mathcal{N}, v) \in \Phi(\mathcal{N}, v)$ *and all symmetric players* $i, j \in \mathcal{N}$ *it holds that* $\phi_i(\mathcal{N}, v) = \phi(\mathcal{N}, v)_j$.

The symmetry property implies that if two features have the same marginal contribution to the goodness of fit when added to any coalition then the importance of the two features is the same. This property is essentially a fair treatment of inputs and results in identical features receiving the same importance score.

Definition 11. ***Linearity.*** *A single-valued solution concept* Φ *satisfies linearity if for any two games* $(\mathcal{N}, v)$ *and* $(\mathcal{N}, w)$, *and for the solution vector of the TU game given by* $(\mathcal{N}, v + w)$ *it holds that*

$$\phi_i(\mathcal{N}, v + w) = \phi_i(\mathcal{N}, v) + \phi_i(\mathcal{N}, w), \quad \forall i \in \mathcal{N}.$$

Let us imagine a binary classifier and two sets of data points – on both of these datasets, we can define feature selection games with binary cross entropy-based payoffs. The Shapley values of input features in the feature selection game calculated on the pooled dataset would be the same as adding together the Shapley values calculated from the two datasets separately.

These four properties together characterize the Shapley value.

Theorem 1 (Shapley, 1953)**.** *A single-valued solution concept satisfies the null player, efficiency, symmetry, and linearity properties if and only if it is the Shapley value.*

3 Approximations of the Shapley Value

Shapley value computation requires a factorial number of characteristic function evaluations, resulting in factorial time complexity. This is prohibitive in a machine learning context when each evaluation can correspond to training a machine learning model.

For this reason, machine learning applications use a variety of Shapley value approximation methods we discuss in this section. In the following discussion $\widehat{\phi}_i^{Sh}$ denotes an approximated Shapley value for player $i \in \mathcal{N}$.

3.1 Monte Carlo Permutation Sampling

Monte Carlo permutation sampling for the general class of cooperative games was first proposed by Castro *et al.* [2009] to approximate the Shapley value in linear time. Their method performs a sampling-based approximation, at each iteration, a random element from the permutations of the player set is drawn. The marginal contributions of the players in the sampled permutation are scaled down by the number of samples (which is equivalent to taking an average) and added to the approximated Shapley values from the previous iteration. Castro *et al.* [2009] provide asymptotic error bounds for this approximation algorithm via the central limit theorem when the variance of the marginal contributions is known. Maleki *et al.* [2013] extended the analysis of this sampling approach by providing error bounds when either the variance or the range of the marginal contributions is known via Chebyshev's and Hoeffding's inequalities. Their bounds hold for a finite number of samples in contrast to the previous asymptotic bounds.

Stratified Sampling for Variance Reduction

In addition to extending the analysis of Monte Carlo estimation, Maleki *et al.* [2013] demonstrate how to improve Shapley value approximation when sampling can be *stratified* by dividing the permutations of the player set into homogeneous, non-overlapping sub-populations. In particular, they show that if the set of permutations can be grouped into strata with similar marginal gains for players, then the approximation will be more precise. Following this, Castro *et al.* [2017] explored stratified sampling approaches using strata defined by the set of all marginal contributions when the player is in a specific position within the coalition. Burgess *et al.* [2021] propose stratified sampling approaches designed to minimize the uncertainty of the estimate via a stratified empirical Bernstein bound.

Other Variance Reduction Techniques

Following the stratified approaches of Maleki *et al.*; Castro *et al.*; Burgess *et al.* [2013; 2017; 2021], Illés *et al.* [2019] propose an alternative variance reduction technique for the sample mean. Instead of generating a random sequence of samples, they instead generate a sequence of ergodic but not independent samples, taking advantage of negative correlation to reduce the sample variance. Mitchell *et al.* [2021] show that other Monte Carlo variance reduction techniques can also be applied to this problem, such as antithetic sampling [Lomeli *et al.*, 2019; Rubinstein *et al.*, 2016]. A simple form of antithetic sampling uses both a randomly sampled permutation and its reverse. Finally, Touati *et al.* [2021] introduce a Bayesian Monte Carlo approach to Shapley value calculation, showing that Shapley value estimation can be improved by using Bayesian methods.

3.2 Multilinear Extension

By inducing a probability distribution over the subsets $\mathcal{S}$ where $\mathcal{E}_i$ is a random subset that does not include player i and each player is included in a subset with probability q, Owen [1972] demonstrated that the sum over subsets in Definition 7 can also be represented as an integral $\int_0^1 e_i(q)dq$ where $e_i(q) = \mathbb{E}[v(\mathcal{E}_i \cup i) - v(\mathcal{E}_i)]$. Sampling over q therefore provides an approximation method – the multilinear extension. For example, Mitchell *et al.* [2021] uses the trapezoid rule to sample q at fixed intervals while Okhrati *et al.* [2021] proposes incorporating antithetic sampling as a variance reduction technique.

3.3 Linear Regression Approximation

In their seminal work Lundberg *et al.* [2017] apply Shapley values to feature importance and explainability (SHAP values), demonstrating that Shapley values for TU games can be approximated by solving a weighted least squares optimization problem. Their main insight is the computation of Shapley values by approximately solving the following optimization problem:

$$w_{\mathcal{S}} = \frac{|\mathcal{N}| - 1}{\binom{|\mathcal{N}|}{|\mathcal{S}|}|\mathcal{S}|(|\mathcal{N}| - |\mathcal{S}|)} \tag{2}$$

$$\underset{\widehat{\phi}_0^{Sh},\ldots,\widehat{\phi}_n^{Sh}}{min} \sum_{\mathcal{S}\subseteq\mathcal{N}} w_{\mathcal{S}} \left(\widehat{\phi}_0^{Sh} + \sum_{i\in\mathcal{S}} \widehat{\phi}_i^{Sh} - v(\mathcal{S}) \right) \tag{3}$$

$$s.t.\ \ \widehat{\phi}_0^{Sh} = v(\emptyset), \qquad \widehat{\phi}_0^{Sh} + \sum_{i\in\mathcal{N}} \widehat{\phi}_i^{Sh} = v(\mathcal{N}). \tag{4}$$

The definition of weights in Equation (2) and the objective function in Equation (3) implies the evaluation of $v(\cdot)$ for 2^n coalitions. To address this Lundberg *et al.* [2017] propose approximating this by subsampling the coalitions. Note that $w_{\mathcal{S}}$ is higher when coalitions are large or small. Covert *et al.* [2021] extend the study of this method, finding that while SHAP is a consistent estimator, it is not an unbiased one. By proposing and analyzing a variation of this unbiased method, they conclude that while there is a small bias incurred by SHAP it has a significantly lower variance than the corresponding unbiased estimator. Covert *et al.* [2021] then propose a variance reduction method for SHAP, improving convergence speed by a magnitude through sampling coalitions in pairs with each selected alongside its complement.

4 Machine Learning and the Shapley Value

Our discussion about applications of the Shapley value in machine learning focuses on the formulation of the cooperative games, definition of the player set and payoffs, Shapley value approximation, and the time complexity of the approximation. We summarized the most important application areas in Table 2 and grouped the relevant works by the problem solved.

4.1 Feature Selection

The feature selection game treats input features of a machine learning model as players and model performance as the payoff [Guyon *et al.*, 2003; Fryer *et al.*, 2021]. The Shapley values of features quantify how much individual features contribute to the model's performance on a set of data points.

Definition 12. ***Feature selection game.*** *Let the player set be $\mathcal{N} = \{1, \ldots, n\}$, for $\mathcal{S} \subseteq \mathcal{N}$ the train and test feature vector sets are $\mathcal{X}_{\mathcal{S}}^{Train} = \{\boldsymbol{x}_i^{Train} | i \in \mathcal{S}\}$ and $\mathcal{X}_{\mathcal{S}}^{Test} = \{\boldsymbol{x}_i^{Test} | i \in \mathcal{S}\}$. Let $f_{\mathcal{S}}(\cdot)$ be a machine learning model trained using $\mathcal{X}_{\mathcal{S}}^{Train}$ as input, then the payoff is $v(\mathcal{S}) = g(\boldsymbol{y}, \widehat{\boldsymbol{y}}_{\mathcal{S}})$ where $g(\cdot)$ is a*

Application	Reference	Payoff	Approximation	Time
Feature Selection	[Cohen *et al.*, 2007]	Validation loss	Exact	$\mathcal{O}(\lvert\mathcal{N}\rvert!)$
	[Sun *et al.*, 2012]	Mutual information	Exact	$O(\lvert\mathcal{N}\rvert!)$
	[Williamson *et al.*, 2020]	Validation loss	Monte Carlo sampling	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Tripathi *et al.*, 2020]	Training loss	Monte Carlo sampling	$O(\lvert\mathcal{N}\rvert)$
	[Patel *et al.*, 2021]	Validation loss	Monte Carlo sampling	$O(\lvert\mathcal{N}\rvert)$
	[Guha *et al.*, 2021]	Validation loss	Exact	$O(\lvert\mathcal{N}\rvert!)$
Data Valuation	[Jia *et al.*, 2019]	Validation loss	Restricted Monte Carlo sampling	$\mathcal{O}(\sqrt{\lvert\mathcal{N}\rvert}\log\lvert\mathcal{N}\rvert^2)$
	[Ghorbani *et al.*, 2019]	Validation loss	Monte Carlo sampling	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Shim *et al.*, 2021]	Validation loss	Exact	$\mathcal{O}(\lvert\mathcal{N}\rvert\log\lvert\mathcal{N}\rvert)$
	[Deutch *et al.*, 2021]	Validation loss	Restricted Monte Carlo sampling	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Kwon *et al.*, 2021a]	Validation loss	Monte Carlo sampling	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Kwon *et al.*, 2021b]	Validation loss	Monte Carlo sampling	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
Federated Learning	[Liu *et al.*, 2021]	Validation loss	Monte Carlo sampling	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
Universal Explainability	[Lundberg *et al.*, 2017]	Attribution	Linear regression	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Sundararajan *et al.*, 2020a]	Interaction attribution	Integrated gradients	$\mathcal{O}(\lvert\mathcal{N}\rvert^2)$
	[Sundararajan *et al.*, 2020b]	Interaction attribution	Integrated gradients	$\mathcal{O}(\lvert\mathcal{N}\rvert^2)$
	[Frye *et al.*, 2020a]	Attribution	Linear regression	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Frye *et al.*, 2020b]	Attribution	Linear regression	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Yuan *et al.*, 2021]	Attribution	Monte Carlo sampling	$O(\lvert\mathcal{N}\rvert)$
	[Covert *et al.*, 2021]	Attribution	Linear regression	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
Explainability of Deep Learning	[Chen *et al.*, 2018]	Attribution	Restricted Monte Carlo sampling	$\mathcal{O}(2^{\lvert\mathcal{N}\rvert})$ or $\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Ancona *et al.*, 2019]	Neuron attribution	Voting game	$\mathcal{O}(\lvert\mathcal{N}\rvert^2)$
	[Ghorbani *et al.*, 2020b]	Neuron attribution	Monte Carlo sampling	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Zhang *et al.*, 2021]	Interaction Attribution	Linear regression	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
Explainability of Graphical Models	[Liu *et al.*, 2020]	Attribution	Exact	$\mathcal{O}(\lvert\mathcal{N}\rvert!)$
	[Heskes *et al.*, 2020]	Causal Attribution	Linear regression	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Wang *et al.*, 2021b]	Causal Attribution	Linear regression	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Singal *et al.*, 2021]	Causal Attribution	Linear regression	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
Explainability in Graph Machine Learning	[Yuan *et al.*, 2021]	Edge level attribution	Monte Carlo sampling	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Duval *et al.*, 2021]	Edge level attribution	Linear regression	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
Multi-agent Reinforcement Learning	[Wang *et al.*, 2021a]	Global reward	Monte Carlo sampling	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
	[Li *et al.*, 2021]	Global reward	Monte Carlo sampling	$\mathcal{O}(\lvert\mathcal{N}\rvert)$
Model Valuation in Ensembles	[Rozemberczki *et al.*, 2021]	Predictive performance	Voting game	$\mathcal{O}(\lvert\mathcal{N}\rvert^2)$

Table 2: An application area, payoff definition, Shapley value approximation technique, and computation time (the player set is noted by $\mathcal{N}$) based comparison of research works. Specific applications of the Shapley value are grouped together and ordered chronologically.

goodness of fit function, $\mathbf{y}$ and $\widehat{\mathbf{y}}_{\mathcal{S}} = f_{\mathcal{S}}(\mathcal{X}_{\mathcal{S}}^{Test})$ are the ground truth and predicted targets.

Shapley values, and close relatives such as the Banzhaf index [Banzhaf III, 1964], have been studied as a measure of feature importance in various contexts [Cohen *et al.*, 2007; Pintér, 2011; Sun *et al.*, 2012; Williamson *et al.*, 2020; Tripathi *et al.*, 2020]. Using these importance estimates, features can be ranked and selected or removed accordingly. This approach has been applied to various tasks such as vocabulary selection in natural language processing [Patel *et al.*, 2021] and feature selection in human action recognition [Guha *et al.*, 2021].

4.2 Data Valuation

In the data valuation game training set data points are players and the payoff is defined by the goodness of fit achieved by a model on the test data. Computing the Shapley value of players in a data valuation game measures how much data points contribute to the performance of the model.

Definition 13. ***Data valuation game.*** *Let the player set be $\mathcal{N} = \{(\boldsymbol{x}_i, y_i) \mid 1 \leq i \leq n\}$ where $\boldsymbol{x}_i$ is the input feature vector and y_i is the target. Given the coalition $\mathcal{S} \subseteq \mathcal{N}$ let $f_{\mathcal{S}}(\cdot)$ be a machine learning model trained on $\mathcal{S}$. Let us denote the test set feature vectors and targets as $\mathcal{X}$ and $\mathcal{Y}$, given $f_{\mathcal{S}}(\cdot)$ the set of predicted labels is defined as $\widehat{\mathcal{Y}} = \{f_{\mathcal{S}}(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{X}\}$. Then the payoff of a model trained on the data points $\mathcal{S} \subseteq \mathcal{N}$ is $v(\mathcal{S}) = g(\mathcal{Y}, \widehat{\mathcal{Y}})$ where $g(\cdot)$ is a goodness of fit metric.*

The Shapley value is not the only method for data valuation – earlier works used function utilization [Koh *et al.*, 2017; **?**], leave-one-out testing [Cook, 1977] and core sets [Dasgupta *et al.*, 2009]. However, these methods fall short [Jia *et al.*, 2019; Ghorbani *et al.*, 2019; Kwon *et al.*, 2021b] when there are fairness requirements from the data valuation technique. Ghorbani *et al.* [2019] proposed a framework of utilizing Shapley value in a data-sharing system; Jia *et al.* [2019] advanced this work with more efficient algorithms to approximate the Shapley value for data valuation. The distributional Shapley value has been discussed by Ghorbani *et al.* [2020a] who argued that maintaining privacy is hard during Shapley value computation. Their method calculates the Shapley value over a distribution which solves problems such as lack of privacy. The computation time of this can be reduced as Kwon *et al.* [2021a] point out with approximation methods optimized for specific models.

4.3 Federated Learning

A federated learning scenario can be seen as a cooperative game by modeling the data owners as players who cooperate to train a high-quality machine learning model [Liu *et al.*, 2021].

Definition 14. ***Federated learning game.*** *In this game players are a set of labeled dataset owners $\mathcal{N} = \{(\mathcal{X}_i, \mathcal{Y}_i) | 1 \leq i \leq n\}$ where $\mathcal{X}_i$ and $\mathcal{Y}_i$ are the feature and label sets owned by the i^{th} silo. Let $(\mathcal{X}, \mathcal{Y})$ be a labeled test set, $\mathcal{S} \subseteq \mathcal{N}$ a coalition of data silos, $f_{\mathcal{S}}(\cdot)$ a machine learning model trained on $\mathcal{S}$, and $\widehat{\mathcal{Y}}_{\mathcal{S}}$ the labels predicted by $f_{\mathcal{S}}(\cdot)$ on $\mathcal{X}$. The payoff of $\mathcal{S} \subseteq \mathcal{N}$ is $v(\mathcal{S}) = g(\mathcal{Y}, \widehat{\mathcal{Y}}_{\mathcal{S}})$ where $g(\cdot)$ is a goodness of fit metric.*

The system described by Liu *et al.* [2021] uses Monte Carlo sampling to approximate the Shapley value of the data silos.

Given the potentially overlapping nature of the datasets, the use of configuration games, an extension of the Shapley value, could be an interesting future direction for federated learning.

4.4 Explainable Machine Learning

In explainable machine learning the Shapley value is used to measure the contributions of input features to the output of a machine learning model at the instance level. Given a specific data point, the goal is to decompose the model prediction and assign Shapley values to individual features of the instance. There are universal solutions to this challenge that are model agnostic and designs customized for deep learning [Chen *et al.*, 2018; Ancona *et al.*, 2019], classification trees [Lundberg *et al.*, 2017], and graphical models [Liu *et al.*, 2020; Singal *et al.*, 2021].

Universal Explainability

A cooperative game for universal explainability is completely model agnostic; the only requirement is that a scalar-valued output can be generated by the model such as the probability of a class label being assigned to an instance.

Definition 15. ***Universal explainability game.*** *Let us denote the machine learning model of interest with $f(\cdot)$ and let the player set be the feature values of a single data instance: $\mathcal{N} = \{x_i | 1 \leq i \leq n\}$. The payoff of a coalition $\mathcal{S} \subseteq \mathcal{N}$ in this game is the scalar valued prediction $v(\mathcal{S}) = \widehat{y}_{\mathcal{S}} = f(\mathcal{S})$ calculated from the subset of feature values.*

Calculating the Shapley value in a game like this offers a complete decomposition of the prediction because the *efficiency* axiom holds. The Shapley values of feature values are explanatory attributions to the input features and missing input feature values are imputed with a reference value such as the mean computed from multiple instances [Lundberg *et al.*, 2017; Covert *et al.*, 2021]. The pioneering Shapley value-based universal explanation method SHAP [Lundberg *et al.*, 2017] proposes a linear time approximation of the Shapley values which we discussed in Section 3. This approximation has shortcomings and implicit assumptions about the features which are addressed by newer Shapley value-based explanation techniques. For example, in [Frye *et al.*, 2020a] the input features are not necessarily independent, [Frye *et al.*, 2020b] restricts the permutations based on known causal relationships, and in [Covert *et al.*, 2021] the proposed technique improves the convergence guarantees of the approximation. Several methods generalize SHAP beyond feature values to give attributions to first-order feature interactions [Sundararajan *et al.*, 2020b; Sundararajan *et al.*, 2020a]. However, this requires that the player set is redefined to include feature interaction values.

Deep Learning

In neuron explainability games neurons are players and attributions to the neurons are payoffs. The primary goal of Shapley value-based explanations in deep learning is to solve these games and compute attributions to individual neurons and filters [Ghorbani *et al.*, 2020b; Ancona *et al.*, 2019].

Definition 16. ***Neuron explainability game.*** *Let us consider $f_{\text{IN}}(\cdot)$ the encoder layer of a neural network and $\boldsymbol{x}$ the input feature vector to the encoder. In the neuron explainability game the player set is $\mathcal{N} = f_{\text{IN}}(\boldsymbol{x}) = \{h_1, \ldots, h_n\}$ - each player corresponds to the output of a neuron in the final layer of the encoder. The payoff of coalition $\mathcal{S} \subseteq \mathcal{N}$ is defined as the predicted output $v(\mathcal{S}) = \widehat{y}_{\mathcal{S}} = f_{\text{OUT}}(\mathcal{S})$ where $f_{\text{OUT}}(\cdot)$ is the head layer of the neural network.*

In practical terms, the payoffs are the output of the neural network obtained by masking out certain neurons. Using the Shapley values obtained in these games the value of individual neurons can be quantified. At the same time, some deep learning specific Shapley value-based explanation techniques have designs and goals that are aligned with the games described in universal explainability. These methods exploit the structure of the input data [Chen *et al.*, 2018] or the nature of feature interactions [Zhang *et al.*, 2021] to provide efficient computations of attributions.

Graphical Models

Compared to universal explanations the graphical model-specific techniques restrict the admissible set of player set permutations considered in the attribution process. These restrictions are defined based on known causal relations and permutations are generated by various search strategies on the graph describing the probabilistic model [Heskes *et al.*, 2020; Liu *et al.*, 2020; Singal *et al.*, 2021]. Methods are differentiated from each other by how restrictions are defined and how permutations are restricted.

Relational Machine Learning

In the relational machine learning domain the Shapley value is used to create edge importance attributions of instance-level explanations [Duval *et al.*, 2021; Yuan *et al.*, 2021]. Essentially the Shapley value in these games measures the average marginal change in the outcome variable as one adds a specific edge to the edge set in all of the possible edges set permutations. It is worth noting that the edge explanation and attribution techniques proposed could be generalized to provide node attributions.

Definition 17. ***Relational explainability game.*** *Let us define a graph $\mathcal{G} = (\mathcal{V}, \mathcal{N})$ where $\mathcal{V}$ and $\mathcal{N}$ are the vertex and edge sets. Given the relational machine learning model $f(\cdot)$, node feature matrix $\boldsymbol{X}$, node $u \in \mathcal{V}$, the payoff of coalition $\mathcal{S} \subseteq \mathcal{V}$ in the graph machine learning explanation game is defined as the node level prediction $v(\mathcal{S}) = \widehat{y}_{\mathcal{S},u} = f(\boldsymbol{X}, \mathcal{V}, \mathcal{S}, u)$.*

4.5 Multi-Agent Reinforcement Learning

Global reward multi-agent reinforcement learning problems can be modeled as TU games [Wang *et al.*, 2021a; Li *et al.*, 2021] by defining the player set as the set of agents and the payoff of coalitions as a global reward. The Shapley value allows an axiomatic decomposition of the global reward achieved by the agents and the fair attribution of credit assignments to agents.

4.6 Model Valuation in Ensembles

The Shapley value can be used to assess the contributions of machine learning models to a composite model in ensemble games. In these games, players are models in an ensemble and payoffs are decided by whether an instance level prediction made by the model is correct.

Definition 18. ***Ensemble game****. Let us consider a single target - feature instance denoted by $(y, \boldsymbol{x})$. The player set in ensemble games is defined by a set of machine learning models $\mathcal{N} = \{f_i(\cdot) | 1 \leq i \leq n\}$ that operate on the feature set. The predicted target output by the ensemble $\mathcal{S} \subseteq \mathcal{N}$ is defined as $\widehat{y}_{\mathcal{S}} = \tilde{f}(\mathcal{S}, \boldsymbol{x})$*

where $\tilde{f}(\cdot)$ is a prediction aggregation function. The payoff of $\mathcal{S}$ is $v(\mathcal{S}) = g(y, \widehat{y}_{\mathcal{S}})$ where $g(\cdot)$ is a goodness of fit metric.

The ensemble games described by [Rozemberczki *et al.*, 2021] are formulated as a special subclass of voting games. This allows the use of precise game-specific approximation [Fatima *et al.*, 2008] techniques and because of this the Shapley value estimates are obtained in quadratic time and have a tight approximation error. The games themselves are model agnostic concerning the player set – ensembles can be formed by heterogeneous types of machine learning models that operate on the same inputs.

5 Discussion

The Shapley value has a wide-reaching impact in machine learning, but it has limitations and certain extensions of the Shapley value could have important applications in machine learning.

5.1 Limitations

Computation Time

Computing the Shapley value for each player naively in a TU game takes factorial time. In some machine learning application areas such as multi-agent reinforcement learning and federated learning where the number of players is small, this is not an issue. However, in large scale data valuation [Kwon *et al.*, 2021a; Kwon *et al.*, 2021b], explainability [Lundberg *et al.*, 2017], and feature selection [Patel *et al.*, 2021] settings the exact calculation of the Shapley value is not tractable. In Sections 3 and 4 we discussed approximation techniques proposed to make Shapley value computation possible. In some cases, asymptotic properties of these Shapley value approximation techniques are not well understood – see for example [Chen *et al.*, 2018].

Interpretability

By definition, the Shapley values are the average marginal contributions of players to the payoff of the grand coalition computed from all permutations [Shapley, 1953]. Theoretical interpretations like this one are not intuitive and not useful for non-game theory experts. This means that translating the meaning of Shapley values obtained in many application areas to actions is troublesome [Kumar *et al.*, 2020]. For example in a data valuation scenario: is a data point with a twice as large Shapley value as another one twice as valuable? Answering a question like this requires a definition of a cooperative game that is intuitive.

Axioms Do Not Hold Under Approximations

As we discussed most applications of the Shapley value in machine learning use approximations. The fact that under these approximations the desired axiomatic properties of the Shapley value do not hold is often overlooked [Sundararajan *et al.*, 2020b]. This is problematic because most works argue for the use of Shapley value based on these axioms. In our view, this is the greatest unresolved issue in the application of the Shapley value.

5.2 Future Research Directions

Hierarchy of the Coalition Structure

The Shapley value has a constrained version called Owen value [Owen, 1977] in which only permutations satisfying conditions defined by a *coalition structure* - a partition of the player set - are considered. The calculation of the Owen value is identical to that of the Shapley value, with the exception that only those permutations are taken into account where the players in any of the subsets of the *coalition structure* follow each other. In several real-world data and feature valuation scenarios even more complex hierarchies of the coalition, the structure could be useful. Having a nested hierarchy imposes restrictions on the admissible permutations of the players and changes player valuation. Games with such nested hierarchies are called level structure games in game theory. [Winter, 1989] presents the Winter value a solution concept to level structure games - such games are yet to receive attention in the machine learning literature.

Overlapping Coalition Structure

Traditionally, it is assumed that players in a coalition structure are allocated in disjoint partitions of the grand coalition. Allowing players to belong to overlapping coalitions in configuration games [Albizuri *et al.*, 2006] could have several applications in machine learning. For example in a data-sharing - feature selection scenario multiple data owners might have access to the same features - a feature can belong to overlapping coalitions.

Solution Concepts Beyond the Shapley Value

The Shapley value is a specific solution concept of cooperative game theory with intuitive axiomatic properties (Section 2). At the same time it has limitations with respect to computation constraints and interpretability (Sections 3 and 5). Cooperative game theory offers other solution concepts such as the *core*, *nucleolus*, *stable set*, and *kernel* with their own axiomatizations. For example, the *core* has been used for model explainability and feature selection [Yan *et al.*, 2021]. Research into the potential applications of these solution concepts is lacking.

6 Conclusion

In this survey we discussed the Shapley value, examined its axiomatic characterizations and frequently used Shapley value approximations. We defined and reviewed its uses in machine learning, highlighted issues with the Shapley value and potential new application and novel research areas in machine learning.

Acknowledgements

This research was supported by REPHRAIN: The National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (UKRI grant: EP/V011189/1). The authors would like to thank Anton Tsitsulin for feedback throughout the preparation of this manuscript.

References

[Albizuri *et al.*, 2006] Josune Albizuri, Jesús Aurrecoechea, et al. Configuration Values: Extensions of the Coalitional Owen Value. *Games and Economic Behavior*, pages 1–17, 2006.

[Ancona *et al.*, 2019] Marco Ancona, Cengiz Oztireli, and Markus Gross. Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Value Approximation. In *International Conference on Machine Learning*, pages 272–281, 2019.

[Banzhaf III, 1964] John F Banzhaf III. Weighted Voting Doesn't Work: A Mathematical Analysis. *Rutgers L. Rev.*, page 317, 1964.

[Burgess *et al.*, 2021] Mark A Burgess and Archie C Chapman. Approximating the Shapley Value Using Stratified Empirical Bernstein Sampling. International Joint Conferences on Artificial Intelligence Organization, 2021.

[Castro *et al.*, 2009] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial Calculation of the Shapley Value Based on Sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[Castro *et al.*, 2017] Javier Castro, Daniel Gómez, et al. Improving Polynomial Estimation of the Shapley Value by Stratified Random Sampling with Optimum Allocation. *Computers & Operations Research*, 82:180–188, 2017.

[Chalkiadakis *et al.*, 2011] Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. Computational Aspects of Cooperative Game Theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(6):1–168, 2011.

[Chen *et al.*, 2018] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. In *International Conference on Learning Representations*, 2018.

[Cohen *et al.*, 2007] Shay Cohen, Gideon Dror, and Eytan Ruppin. Feature Selection via Coalitional Game Theory. *Neural Computation*, (7):1939–1961, 2007.

[Cook, 1977] R Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 1977.

[Covert *et al.*, 2021] Ian Covert and Su-In Lee. Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3457–3465, 2021.

[Dasgupta *et al.*, 2009] Anirban Dasgupta, Petros Drineas, , et al. Sampling algorithms and coresets for \ell_p regression. *SIAM Journal on Computing*, pages 2060–2078, 2009.

[Deutch *et al.*, 2021] Daniel Deutch, Nave Frost, Amir Gilad, and Oren Sheffer. *Explanations for Data Repair Through Shapley Values*, page 362–371. 2021.

[Duval *et al.*, 2021] Alexandre Duval and Fragkiskos D. Malliaros. Graphsvx: Shapley value explanations for graph neural networks. In *Machine Learning and Knowledge Discovery in Databases.*, pages 302–318, 2021.

[Fan *et al.*, 2021] Zhenan Fan, Huang Fang, Zirui Zhou, et al. Improving Fairness for Data Valuation in Federated Learning. *arXivv:2109.09046*, 2021.

[Fatima *et al.*, 2008] Shaheen S Fatima, Michael Wooldridge, and Nicholas R Jennings. A Linear Approximation Method for the Shapley Value. *Artificial Intelligence*, 172(14):1673–1699, 2008.

[Frye *et al.*, 2020a] Christopher Frye, Damien de Mijolla, Tom Begley, et al. Shapley Explainability on the Data Manifold. In *International Conference on Learning Representations*, 2020.

[Frye *et al.*, 2020b] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric Shapley Values: Incorporating Causal Knowledge Into Model-Agnostic Explainability. *Advances in Neural Information Processing Systems*, 33, 2020.

[Fryer *et al.*, 2021] Daniel Fryer, Inga Strümke, and Hien Nguyen. Shapley Values for Feature Selection: the Good, the Bad, and the Axioms. *arXiv:2102.10936*, 2021.

[Ghorbani *et al.*, 2019] Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In *International Conference on Machine Learning*, pages 2242–2251, 2019.

[Ghorbani *et al.*, 2020a] Amirata Ghorbani, Michael Kim, and James Zou. A Distributional Framework for Data Valuation. In *International Conference on Machine Learning*, pages 3535–3544, 2020.

[Ghorbani *et al.*, 2020b] Amirata Ghorbani and James Zou. Neuron Shapley: Discovering the Responsible Neurons. In *Advances in Neural Information Processing Systems*, pages 5922–5932, 2020.

[Guha *et al.*, 2021] Ritam Guha, Ali Hussain Khan, et al. Cga: A new feature selection model for visual human action recognition. *Neural Computing and Applications*, 33(10):5267–5286, 2021.

[Guyon *et al.*, 2003] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[Heskes *et al.*, 2020] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models. *Advances in Neural Information Processing Systems*, 33, 2020.

[Illés *et al.*, 2019] Ferenc Illés and Péter Kerényi. Estimation of the Shapley Value by Ergodic Sampling. *arXiv:1906.05224*, 2019.

[Jia *et al.*, 2019] Ruoxi Jia, David Dao, Boxin Wang, Hubis, et al. Towards Efficient Data Valuation Based on the Shapley Value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176, 2019.

[Koh *et al.*, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894, 2017.

[Kumar *et al.*, 2020] Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-Value-Based Explanations as Feature Importance Measures. In *International Conference on Machine Learning*, pages 5491–5500, 2020.

[Kwon *et al.*, 2021a] Yongchan Kwon, Manuel A Rivas, and James Zou. Efficient Computation and Analysis of Distributional Shapley Values. In *International Conference on Artificial Intelligence and Statistics*, pages 793–801, 2021.

[Kwon *et al.*, 2021b] Yongchan Kwon and James Zou. Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning. *arXiv:2110.14049*, 2021.

[Li *et al.*, 2021] Jiahui Li, Kun Kuang, Baoxiang Wang, et al. Shapley Counterfactual Credits for Multi-Agent Reinforcement Learning. In *Proceedings of the 27th SIGKDD Conference on Knowledge Discovery & Data Mining*, page 934–942, 2021.

[Liu *et al.*, 2020] Yifei Liu, Chao Chen, Yazheng Liu, et al. Shapley Values and Meta-Explanations for Probabilistic Graphical Model Inference. In *Proceedings of the 29th International Conference on Information & Knowledge Management*, pages 945–954, 2020.

[Liu *et al.*, 2021] Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. GTG-Shapley: Efficient and Accurate Participant Contribution Evaluation in Federated Learning. *arXiv:2109.02053*, 2021.

[Lomeli *et al.*, 2019] Maria Lomeli, Mark Rowland, Arthur Gretton, and Zoubin Ghahramani. Antithetic and Monte Carlo Kernel Estimators for Partial Rankings. *Statistics and Computing*, pages 1127–1147, 2019.

[Lundberg *et al.*, 2017] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

[Maleki *et al.*, 2013] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the Estimation Error of Sampling-based Shapley Value Approximation. *arXiv:1306.4265*, 2013.

[Mitchell *et al.*, 2021] Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling Permutations for Shapley Value Estimation. *arXiv:2104.12199*, 2021.

[Okhrati *et al.*, 2021] Ramin Okhrati and Aldo Lipani. A Multilinear Sampling Algorithm to Estimate Shapley Values. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7992–7999. IEEE, 2021.

[Owen, 1972] Guillermo Owen. Multilinear Extensions of Games. *Management Science*, 18(5-part-2):64–79, 1972.

[Owen, 1977] Guilliermo Owen. Values of Games with a Priori Unions. In *Mathematical Economics and Game Theory*, pages 76–88. 1977.

[Patel *et al.*, 2021] Roma Patel, Marta Garnelo, Ian Gemp, et al. Game-Theoretic Vocabulary Selection via the Shapley Value and Banzhaf Index. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2789–2798, 2021.

[Pintér, 2011] Miklós Pintér. Regression games. *Annals of Operations Research*, 186(1):263–274, 2011.

[Rozemberczki *et al.*, 2021] Benedek Rozemberczki and Rik Sarkar. The Shapley Value of Classifiers in Ensemble Games. In *Proceedings of the 30th International Conference on Information and Knowledge Management*, page 1558–1567, 2021.

[Rubinstein *et al.*, 2016] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo Method*. 2016.

[Shapley, 1953] Lloyd Shapley. A Value for N-Person Games. *Contributions to the Theory of Games*, pages 307–317, 1953.

[Shim *et al.*, 2021] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, et al. Online Class-Incremental Continual Learning with Adversarial Shapley Value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9630–9638, 2021.

[Singal *et al.*, 2021] Raghav Singal, George Michailidis, and Hoiyi Ng. Flow-based Attribution in Graphical Models: A Recursive Shapley Approach. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 9733–9743, 2021.

[Sun *et al.*, 2012] Xin Sun, Yanheng Liu, Jin Li, et al. Feature Evaluation and Selection with Cooperative Game Theory. *Pattern recognition*, 45(8):2992–3002, 2012.

[Sundararajan *et al.*, 2020a] Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor Interaction Index. In *International Conference on Machine Learning*, pages 9259–9268, 2020.

[Sundararajan *et al.*, 2020b] Mukund Sundararajan and Amir Najmi. The Many Shapley Values for Model Explanation. In *International Conference on Machine Learning*, pages 9269–9278, 2020.

[Touati *et al.*, 2021] Sofiane Touati, Mohammed Said Radjef, and SAIS Lakhdar. A Bayesian Monte Carlo Method for Computing the Shapley Value: Application to Weighted Voting and Bin Packing Games. *Computers & Operations Research*, 125:105094, 2021.

[Tripathi *et al.*, 2020] Sandhya Tripathi, N Hemachandra, and Prashant Trivedi. Interpretable Feature Subset Selection: A Shapley Value Based Approach. In *IEEE International Conference on Big Data*, pages 5463–5472, 2020.

[Wang *et al.*, 2020] Tianhao Wang, Johannes Rausch, Ce Zhang, et al. A Principled Approach to Data Valuation for Federated Learning. In *Federated Learning*, pages 153–167. 2020.

[Wang *et al.*, 2021a] Jianhong Wang, Jinxin Wang, Yuan Zhang, Yunjie Gu, and Tae-Kyun Kim. SHAQ: Incorporating Shapley Value Theory into Q-Learning for Multi-Agent Reinforcement Learning. *arXiv:2105.15013*, 2021.

[Wang *et al.*, 2021b] Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. Shapley Flow: A Graph-Based Approach to Interpreting Model Predictions. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729, 2021.

[Williamson *et al.*, 2020] Brian Williamson and Jean Feng. Efficient Nonparametric Statistical Inference on Population Feature Importance Using Shapley Values. In *International Conference on Machine Learning*, pages 10282–10291, 2020.

[Winter, 1989] Eyal Winter. A Value for Cooperative Games with Levels Structure of Cooperation. *International Journal of Game Theory*, 18(2):227–40, 1989.

[Yan *et al.*, 2021] Tom Yan and Ariel Procaccia. If You Like Shapley Then You Will Love the Core. *Proceedings of the AAAI Conference*, pages 5751–5759, 2021.

[Yuan *et al.*, 2021] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On Explainability of Graph Neural Networks via Subgraph Explorations. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12241–12252, 2021.

[Zhang *et al.*, 2021] Hao Zhang, Yichen Xie, Longjie Zheng, et al. Interpreting Multivariate Shapley Interactions in DNNs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10877–10886, 2021.