The Shapley Value in Machine Learning

Citation for published version:
Rozemberczki, B, Watson, L, Bayer, P, Yang, H-T, Kiss, O, Nilsson, S & Sarkar, R 2022, The Shapley Value in Machine Learning. in L De Raedt (ed.), Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI-ECAI 2022. International Joint Conferences on Artificial Intelligence Organization, pp. 5572-5579, The 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence, Vienna, Austria, 23/07/22. https://doi.org/10.24963/ijcai.2022/778

Digital Object Identifier (DOI):
10.24963/ijcai.2022/778

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI-ECAI 2022

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
The Shapley Value in Machine Learning

Benedek Rozemberczki\textsuperscript{1}, Lauren Watson\textsuperscript{2}, Péter Bayer\textsuperscript{3}, Hao-Tsung Yang\textsuperscript{2}, Oliver Kiss\textsuperscript{4}, Sebastian Nilsson\textsuperscript{3} and Rik Sarkar\textsuperscript{2}
\textsuperscript{1}Research Data & Analytics, Research & Development IT, AstraZeneca
\textsuperscript{2}The University of Edinburgh, School of Informatics
\textsuperscript{3} Toulouse School of Economics & Institute for Advanced Study in Toulouse
\textsuperscript{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.\textsuperscript{1}

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 certain 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 of this type is the Shapley value [Shapley, 1953] characterized by several 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.](image)

\textbf{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.

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 \( N = \{1, \ldots, n\} \) be the finite set of players. We call each non-empty subset \( S \subseteq N \) a coalition and \( N \) itself the grand coalition.

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

\textsuperscript{1}The survey is supported by a collection of related work under https://github.com/AstraZeneca/awesome-shapley-value.
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 \( N = \{1, 2, 3\} \). The characteristic function defines the payoff for each coalition. Let these payoffs be given as:

\[
\begin{align*}
v(\emptyset) &= 0; & v(\{1\}) &= 7; & v(\{2\}) &= 11; & v(\{3\}) &= 14; \\
v(\{1, 2\}) &= 18; & v(\{1, 3\}) &= 21; & v(\{2, 3\}) &= 23; & v(\{1, 2, 3\}) &= 25.
\end{align*}
\]

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

Definition 4. Solution concept and solution vector. Solution concept \( \Phi \) is a mapping associating a subset \( \Phi(N, v) \subseteq Z(N, v) \) to every TU game \((N, v)\). A solution vector \( \phi(N, v) \in \mathbb{R}^N \) to the cooperative game \((N, v)\) satisfies solution concept \( \Phi \) if \( \phi(N, v) \in \Phi(N, v) \). Solution concept \( \Phi \) is single-valued if for every \((N, v)\) the set \( \Phi(N, v) \) 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 \( v(N) \). Solution vectors are specific allocations satisfying the principles of the solution concept.

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

Definition 6. Predecessor set. Let the set of predecessors of player \( i \in N \) in permutation \( \pi \) be the coalition:

\[
P^\pi_i = \{ j \in N \mid \pi(j) < \pi(i) \}.
\]

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 1st player is \( P^\pi_1 = \{3, 2\} \), that of the 2nd player is \( P^\pi_2 = \{3\} \) and \( P^\pi_3 = \emptyset \).

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 \((N, v)\) is given by Equation 1.

\[
\phi_{i}^{Sh} = \frac{1}{|H(N)|} \sum_{\pi \in \Pi(N)} [v(P^\pi_i \cup \{i\}) - v(P^\pi_i)]
\]

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.

| Table 1: The permutations of the player set, marginal contributions of the players in each permutation and the Shapley values. |
|-----------------|-----------------|-----------------|-----------------|
| 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)       | 11          | 11          | 12          |
| (3, 1, 2)       | 7           | 4           | 14          |
| (3, 2, 1)       | 2           | 9           | 14          |
| Shapley value   | 32/6        | 50/6        | 68/6        |

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

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 \( \Phi \) is efficient or Pareto optimal if for every game \((N, v)\) and every solution vector \( \phi(N, v) \in \Phi(N, v) \) it holds that \( \sum_{i \in N} \phi(N, v)_i = v(N) \).

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

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

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

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

\[
\phi(N, v + w)_i = \phi(N, v)_i + \phi(N, w)_i, \ \forall i \in N.
\]

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 an exponential number of characteristic function evaluations, resulting in exponential 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 \( \hat{\phi}_i^S \) denotes an approximated Shapley value for player \( i \in N \).

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.

Algorithm 1: Monte Carlo permutation sampling approximation of the Shapley value.

As shown in Algorithm 1, the 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 the 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 to approximate the Shapley value.

3.2 Multilinear Extension

By inducing a probability distribution over the subsets \( S \) 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:

\[
\begin{align*}
\min_{\phi^S_0, \ldots, \phi^S_n} & \quad \sum_{S \subseteq N} w_S \left( \sum_{i \in S} \phi^S_i - v(S) \right) \\
\text{s.t.} & \quad \phi^S_0 = v(\emptyset), \quad \phi^S_n = v(N).
\end{align*}
\]  

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 problem subsampling the coalitions. Note that \( w_S \) 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 estimator. By proposing and analyzing a variation of this method that is unbiased, 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 the machine learning domain focuses on the formulation of the cooperative games, definition of the player set and payoffs, Shapley value approximation technique used, and the time complexity of the approximation. We summarized the most important application areas with this information 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 \( N = \{1, \ldots, n\} \), for \( S \subseteq N \) the train and test feature vectors are \( X_S^{\text{Train}} = \{x_i^{\text{Train}} | i \in S\} \) and \( X_S^{\text{Test}} = \{x_i^{\text{Test}} | i \in S\} \). Let \( f_S(\cdot) \) be a machine learning model trained using \( X_S^{\text{Train}} \) as input, then the payoff is \( v(S) = g(y, \hat{y}_S) \) where \( g(\cdot) \) is a goodness of fit function, \( y \) and \( \hat{y}_S = f_S(X_S^{\text{Test}}) \) 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; 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 \( N = \{(x_i, y_i) | 1 \leq i \leq n\} \) where \( x_i \) is the input feature vector and \( y_i \) is the target. Given the coalition \( S \subseteq N \) let \( f_S(\cdot) \) be a machine learning model trained on \( S \). Let us denote the test set feature vectors and targets as \( X \) and \( Y \), given \( f_S(\cdot) \) the set of predicted labels is defined as \( \hat{Y} = \{f_S(x)|x \in X\} \). Then the payoff of a model trained on the data points \( S \subseteq N \) is \( v(S) = g(Y, \hat{Y}) \) 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; Sharchilev et al., 2018], leave-one-out testing [Cook, 1977] and core sets [Dasgupta et al., 2009]. However, these methods fall short when there are fairness requirements from the data valuation technique [Jia et al., 2019; Ghorbani et al., 2019; Kwon et al., 2021b]. Ghorbani proposed a framework of utilizing Shapley value in a data-sharing system [Ghorbani et al., 2019; 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 keeping 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 machine learning 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 \( \mathcal{N} = \{(X_i, Y_i)|1 \leq i \leq n\} \) where \( X_i \) and \( Y_i \) are the feature and label sets owned by the \( i \)th silo. Let \((X, Y)\) be a labeled test set, \( S \subseteq \mathcal{N} \) a coalition of data silos, \( f_S(\cdot) \) a machine learning model trained on \( S \), and \( \hat{Y}_S \) the labels predicted by \( f_S(\cdot) \) on \( X \). The payoff of \( S \subseteq \mathcal{N} \) is \( v(S) = g(Y, \hat{Y}_S) \) 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 data coming from the data silos in linear time. Given the potentially overlapping nature of the datasets, the use of configuration games could be an interesting future direction [Albizuri et al., 2006].

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: \( N = \{x_i | 1 \leq i \leq n\} \). The payoff of a coalition \( S \subseteq N \) in this game is the scalar valued prediction \( v(S) = \hat{y}_S = f(S) \) 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
Table 2: An application area, payoff definition, Shapley value approximation technique, and computation time (the player set is noted by $N$) based comparison of research works. Specific applications of the Shapley value are grouped together and ordered chronologically.

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

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 $x$ the input feature vector to the encoder. In the neuron explainability game the player set is $N = f_{\text{IN}}(x) = \{h_1, \ldots, h_n\}$ - each player corresponds to the output of a neuron in the final layer of the encoder. The payoff of coalition $S \subseteq N$ is defined as the predicted output $v(S) = \vec{y}_S = f_{\text{OUT}}(S)$ 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 $G = (V, N)$ where $V$ and $N$ are the vertex and edge sets. Given the relational machine learning model $f(\cdot)$, node feature matrix $X$, node $u \in V$, the payoff of coalition $S \subseteq V$ in
The graph machine learning explanation game is defined as the node level prediction \( v(\mathcal{S}) = \hat{g}_{\mathcal{S},u} = f(X, \mathcal{V}, \mathcal{S}, u) \).

### 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 in these games and the fair attribution of credit assignments to each of the participating 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 prediction mad by the model are correct.

**Definition 18. Ensemble game.** Let us consider a single target - feature instance denoted by \((y, x)\). 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 \(\hat{y}_\mathcal{S} = \hat{f}(\mathcal{S}, x)\) where \(\hat{f}(\cdot)\) is a prediction aggregation function. The payoff of \(\mathcal{S}\) is \(v(\mathcal{S}) = g(y, \hat{y}_\mathcal{S})\) where \(g(\cdot)\) is a goodness of fit metric.

The ensemble games described by [Rozenberczki 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 the cooperative game that is interpretable.

**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 applications 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, kernel, stable set, and nucleolus 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 the most frequently used Shapley value approximation approaches. We defined and reviewed its
uses in machine learning, highlighted issues with the Shapley value and potential new application and 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, and José Manuel Zarzuelo. Configuration Values: Extensions of the Coalitional Owen Value. *Games and Economic Behavior*, 57(1):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.*, 19:317, 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, Elisenda Molina, and Juan Tejada. 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*, 19(7):1939–1961, 2007.

[Cook, 1977] R Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 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, Boulos Harb, Ravi Kumar, and Michael W Mahoney. Sampling algorithms and coresets for εl_p regression. *SIAM Journal on Computing*, 38(5):2060–2078, 2009.

[Deuth et al., 2021] Daniel Deuth, Nave Frost, Amir Gilad, and Oren Sheffet. *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, Zaiyi Zhou, Jian Pei, et al. Improving Fairness for Data Valuation in Federated Learning. *arXiv preprint arXiv: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 Migolla, 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 Strüimke, and Hien Nguyen. Shapley Values for Feature Selection: the Good, the Bad, and the Axioms. *arXiv preprint 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, Pawan Kumar Singh, Ram Sarkar, and Debotosh Bhattacharjee. 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 preprint 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 preprint 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 ACM SIGKDD Conference on Knowledge Discovery & Data Mining, page 934–942, 2021.

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

[Liu et al., 2021] Zelie Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. GTG-Shapley: Efficient and Accurate Participant Contribution Evaluation in Federated Learning. arXiv preprint 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, 29(5):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 preprint arXiv:1306.4265, 2013.

[Mitchell et al., 2021] Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling Permutations for Shapley Value Estimation. arXiv preprint 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] Guillermo 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 Garnejo, Ian Gemp, Chris Dyer, and Yoram Bachrach. 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.

[Rozemberczki et al., 2021] Benedek Rozemberczki and Rik Sarkar. The Shapley Value of Classifiers in Ensemble Games. In Proceedings of the 30th ACM 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, volume 10. 2016.

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

[Sharchilev et al., 2018] Boris Sharchilev, Yury Ustinovskiy, Pavel Serdyukov, and Maarten Rijke. Finding influential training samples for gradient boosted decision trees. In International Conference on Machine Learning, pages 4577–4585, 2018.

[Shim et al., 2021] Dongsuh Shim, Zheda Mai, Jiwhan 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 Michalidis, 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, Yancheng Liu, Jin Li, Jianqi Zhu, et al. Feature Evaluation and Selection with Cooperative Game Theory. Pattern recognition, 45(8):3092–3002, 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. Interpretatable 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, Ruoxi Jia, and Dawn Song. A Principled Approach to Data Valuation for Federated Learning. In Federated Learning, pages 153–167, 2020.

[Wang et al., 2021a] Jianhong Wang, Jixin Wang, Yuan Zhang, Yunjie Gu, and Tae-Kyun Kim. SHAQ: Incorporating Shapley Value Theory into Q-Learning for Multi-Agent Reinforcement Learning. arXiv preprint 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 D. Procaccia. If You Like Shapley Then You Will Love the Core. Proceedings of the AAAI Conference on Artificial Intelligence, 35(6):5751–5759, May 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, Die Zhang, and Quanshi Zhang. Interpreting Multivariate Shapley Contributions to the Theory of Games. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 10877–10886, 2021.