


REVIEW

Open Access



Shapley value: from cooperative game to explainable artificial intelligence

Meng Li¹ , Hengyang Sun², Yanjun Huang^{2,3*} and Hong Chen^{1,3}

Abstract

With the tremendous success of machine learning (ML), concerns about their black-box nature have grown. The issue of interpretability affects trust in ML systems and raises ethical concerns such as algorithmic bias. In recent years, the feature attribution explanation method based on Shapley value has become the mainstream explainable artificial intelligence approach for explaining ML models. This paper provides a comprehensive overview of Shapley value-based attribution methods. We begin by outlining the foundational theory of Shapley value rooted in cooperative game theory and discussing its desirable properties. To enhance comprehension and aid in identifying relevant algorithms, we propose a comprehensive classification framework for existing Shapley value-based feature attribution methods from three dimensions: Shapley value type, feature replacement method, and approximation method. Furthermore, we emphasize the practical application of the Shapley value at different stages of ML model development, encompassing pre-modeling, modeling, and post-modeling phases. Finally, this work summarizes the limitations associated with the Shapley value and discusses potential directions for future research.

Keywords: Machine learning, Explainable artificial intelligence, Cooperative game, Shapley value

1 Introduction

Machine learning (ML) has demonstrated great potential in various fields, including games (such as Go [1, 2] and Starcraft [3]), autonomous driving [4, 5], protein structure prediction [6], and natural language processing [7]. However, the growing number of parameters and the increasing complexity of ML models have rendered them increasingly challenging to comprehend, giving rise to concerns. On the one hand, ML models can inadvertently generate unintentional discrimination, which may manifest as biases related to sensitive information such as gender, race, and sexual orientation [8]. On the other hand, legal regulations such as the General Data Protection Regulation (GDPR) of the European Union impose restrictions on personal information and sensitive data, aiming to protect the rights of data subjects [9].

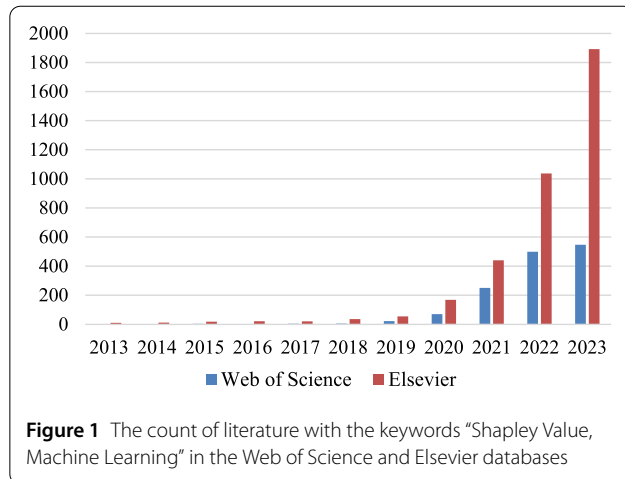
Explainable artificial intelligence (XAI) is a practical approach to unveil the black-box nature of ML models. In the domain of XAI research, local attribution explanations have gained prominence. Several studies have put forth methodologies for local attribution explanations, including Integrated Gradients (IG) [10], Layer-wise Relevance Propagation (LRP) [11], Deep Learning Important Features (DeepLIFT) [12], and Local Interpretable Model-Agnostic Explanations (LIME) [13]. IG calculates the contribution of each feature by considering the integral of gradients along the path from a reference input to the actual input concerning the prediction. LRP propagates the relevance of predictions backward to input features using a backpropagation-like algorithm, thereby comprehensively understanding how the model arrives at its predictions by decomposing it into contributions from each layer. DeepLIFT assigns feature importance by comparing the activation of each feature in the input with a reference activation. LIME fits a simple interpretable model around a prediction sample point and explains the prediction by perturbing input features and observing the re-

* Correspondence: yanjun_huang@tongji.edu.cn

²College of Automotive Studies, Tongji University, Shanghai, 201804, China

³Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai, 201203, China

Full list of author information is available at the end of the article



sulting changes. In 2017, Lundberg and Lee [14] established the Shapley Addition explanation (SHAP), a unified framework for feature attribution measurement based on Shapley value, which integrates several popular feature attribution methods. Shapley value has been proven to be the unique allocation rule that satisfies four axioms: 1) *Efficiency*, which ensures that the sum of the attributions equals the prediction; 2) *Symmetry*, which guarantees that if two features have the same contribution to all possible coalitions, their attributions should be the same; 3) *Dummy player*, which states that a feature that does not affect the prediction should have attribution of zero; and 4) *Additivity/Linearity*, which means that the Shapley value of a player in a game composed of two separate games should be the sum of the Shapley values of the player in each game. Due to its desirable properties, Shapley value has gained popularity as an attribution method. Over the past decade, various Shapley value-based feature attribution methods and applications have also been proposed, as shown in Fig. 1. However, this has also led to the potential for confusion and unintentional misuse of Shapley value-based attribution methods.

To clarify the Shapley value in ML, Rozemberczki *et al.* [15] and Chen *et al.* [16] attempted to summarize the technical aspects of Shapley value feature attribution. Rozemberczki *et al.* [15] discussed the fundamental concepts of cooperative game theory and the axiomatic properties of Shapley value, followed by an overview of Shapley value’s applications in ML. Chen *et al.* [16] primarily focused on estimation algorithms of Shapley value-based feature attribution. It delves into the intricacies of estimating feature attribution and offers insights into various methodologies and estimation strategies. In contrast, our study provides a more detailed classification of algorithms related to Shapley value and a comprehensive review of its applications and limitations. Specifically, our investigation begins with thoroughly examining the fundamental principles of Shapley value in cooperative game theory. Subse-

quently, we present a comprehensive categorization of current feature attribution techniques rooted in Shapley values, elucidating their practical implementations, inherent constraints, and potential directions for further research.

The main contributions are fourfold:

- This paper presents a clear introduction to the origins of Shapley values in ML, specifically focusing on providing a comprehensive understanding of Shapley values and restricted Shapley values derived from cooperative game theory.
- A three-dimensional classification framework is proposed, categorizing existing Shapley value-based feature attribution methods into the dimensions of Shapley value type, feature replacement method, and approximation method. This framework enables researchers to identify relevant algorithms within the field effectively.
- Detailed applications of Shapley values are provided in the three stages of ML model development: pre-modeling, in-modeling, and post-modeling. This comprehensive coverage enhances the practical understanding of how Shapley values can be utilized throughout the ML pipeline.
- This study offers a comprehensive understanding of the limitations of the Shapley value algorithm and provides insights for future research directions, which assists researchers in identifying algorithmic deficiencies and inspires intriguing studies in the field.

The remaining sections of this paper are organized as follows: Sect. 2 provides a comprehensive review of cooperative game theory, focusing on the foundational theory of Shapley value and its various variants. Section 3 introduces the application of Shapley values in the context of ML and explores the different approximation algorithms used for calculating Shapley values. Section 4 delves into the constrained Shapley value. Section 5 presents a thorough examination of the applications of Shapley values in ML, covering their usage across the pre-modeling, modeling, and post-modeling stages. Section 6 discusses the limitations of estimation methods for Shapley value and provides suggestions for future research. Finally, Sect. 7 offers a concise summary of the key findings presented in this paper.

2 Shapley value and Owen value in cooperative games

This section reviews the fundamental theory of cooperative games and then presents the Shapley value and the restricted Shapley value (Owen value) with coalition structures in cooperative games.

2.1 Shapley value in cooperative games

A Transferable Utility (TU) cooperative game [18], often referred to as a TU game, is represented by a binary tuple

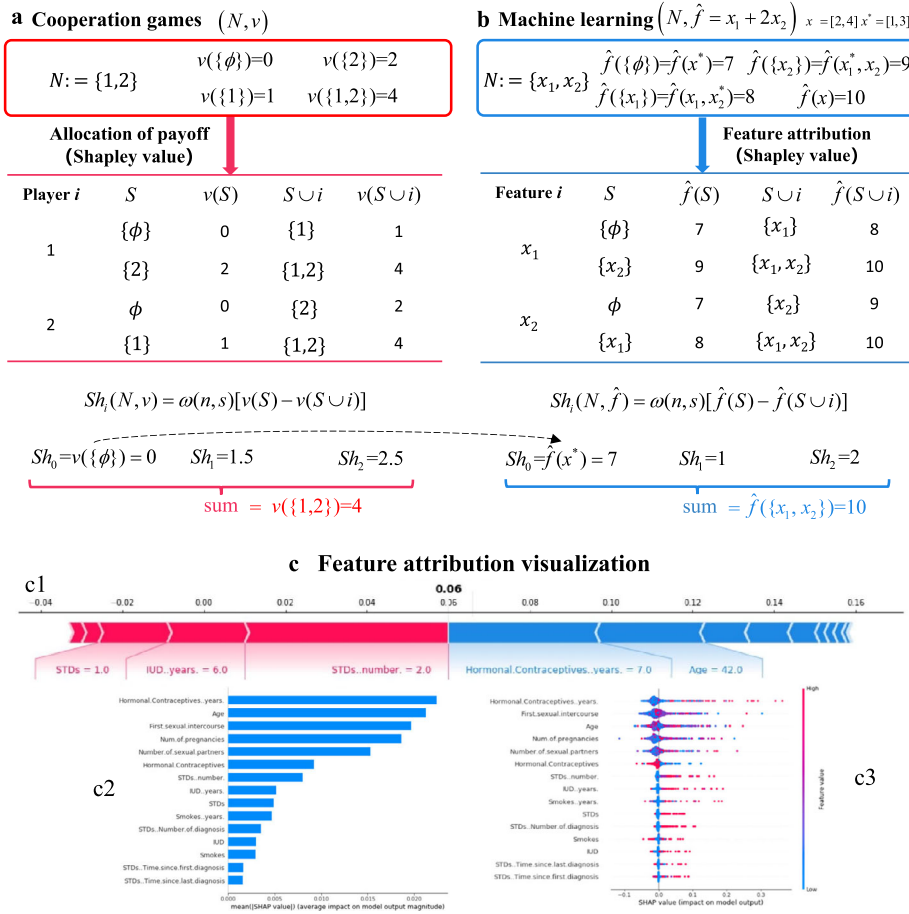


Figure 2 Shapley value computation: from cooperative game to ML. **a**, a general process for allocating contributions using the Shapley value in a cooperative game, where the value $v(S)$ for each subset S is determined, and the allocation Sh_i of contributions for each participant i is then calculated based on (1). **b**, a general process for feature attribution using the Shapley value in an ML model, where the features not included in S are replaced with their corresponding features from the background data x^* , and subsequently, $\hat{f}(S)$ is computed, and the allocation Sh_i of feature attribution for each feature i is then calculated based on (2). It is worth noting that the Shapley value of the empty set ϕ is 0 in cooperative games, while in ML models, the Shapley value is $\hat{f}(x^*)$. **c**, The visualization of Shapley value-based feature attributions. c1. Local explanation: For a machine learning-based cervical cancer risk prediction model, a certain test individual is predicted to have a low risk, at 0.06. The local explanation shows that the impact of increased risk factors (represented by red bars indicating positive contributions) such as sexually transmitted diseases (STDs) is offset by the effect of decreased risk factors (represented by blue bars indicating negative contributions) like age. c2. Feature Importance: Measured by the absolute average of Shapley values, where the number of years using hormonal contraceptives is the most important feature, causing an average change of 2.4 percentage points in the absolute predicted cancer probability (x-axis at 0.024). c3. Shapley Value Summary Plot: Fewer years of using hormonal contraceptives will reduce the predicted cancer risk, while more years will increase the risk [14, 17]

(N, v) . In this representation, the set of players is denoted as $N = \{1, 2, \dots, n\}$. The characteristic function v is a mapping: $2^N \rightarrow \mathbb{R}$ with $v(\phi) = 0$, representing an empty coalition ϕ with zero value. For any subset $S \subseteq N$, the term $v(S)$ denotes the value generated by the collaborative efforts of participants in coalition S . The allocation rule $\phi_i(N, v)$ is the distribution of contributions to participant i :

$$\phi_i(N, v) = Sh_i(N, v) = \underbrace{\omega(n, s)}_{\text{weight factor}} \underbrace{(v(S \cup \{i\}) - v(S))}_{\text{marginal contribution}}, \quad (1)$$

where $\omega(n, s) = \sum_{S \subseteq 2^N, i \notin S} \frac{s!(n-s-1)!}{n!}$, s denotes the number of members in alliance S . The Shapley value [19] has been mathematically proven to be the unique allocation that satisfies the following four axioms.

- *Efficiency*: The sum of the payoffs allocated to all players equals the total value of the game: $\sum_{i=1}^n \phi_i(N, v) = v(N)$.
- *Symmetry*: If two players i and j contribute equally to all possible coalitions, i.e., for any coalition S , $v(S \cup \{i\}) - v(S) = v(S \cup \{j\}) - v(S)$, then their Shapley values must be equal, $\phi_i(N, v) = \phi_j(N, v)$ for all $i, j \in N$.

- *Dummy Player*: If a player i does not contribute to any coalition, i.e., for any coalition S , $v(S \cup \{i\}) = v(S)$, then their Shapley value should be zero, $\phi_i(N, v) = 0$.
- *Additivity/Linearity*: For two cooperative games (N, v_1) and (N, v_2) , the Shapley value of the combined game $(N, v_1 + v_2)$ with characteristic function $v_1 + v_2$ is the sum of the Shapley values of each game, $\phi_i(N, v_1 + v_2) = \phi_i(N, v_1) + \phi_i(N, v_2)$.

A motivation example is shown in Fig. 2(a), illustrating a general process for allocating contributions using the Shapley value in a cooperative game. It is crucial to note that the value function for the empty set is assigned a value of 0, and the sum of the contributions of all participants is equal to the $v(N)$.

2.2 Owen value in cooperative games

The Owen value is an extension of the Shapley value for games with a priori unions or coalitions. Shapley values consider each participant in a coalition as an individual without considering the coalition structure. As some participants may prefer to act together as sub-coalitions, we use a partition denoted by $C = \{\{C_1\}, \{C_2\}, \dots, \{C_m\}\}$ to represent the coalition structure of a set N . Each $C_k \in C$ with $k \in M$ is a priority coalition. Specifically, we define $M = \{1, 2, \dots, m\}$ as the set of indices for the priority coalitions. Additionally, define $C^n = \{\{1\}, \{2\}, \dots, \{n\}\}$ as the trivial coalition structure, where each forms a separate coalition. A triplet (N, v, c) represents a TU game with a coalition structure. In this context, one restricted form of Shapley values, Owen value [20], is given:

$$\begin{aligned} \phi_i(N, v, C) &= Ow_i(N, v, C) \\ &= \sum_{R \subseteq M, (k) T \in S_k \setminus \{i\}} \frac{(|M| - |R| - 1)! |R|! (|S_k| - |T| - 1)! |T|!}{|M|!} \\ &\quad \cdot [v(R \cup T \cup \{i\}) - v(R \cup T)], \end{aligned} \quad (2)$$

where $|R|$ refers to the number of sub-coalitions before S_k , while $|M| - |R| - 1$ represents the number of sub-coalitions after S_k . The term $|T|$ denotes the count of other participants within the sub-coalition S_k appearing before participant i , and likewise, $|S_k| - |T| - 1$ represents the count of other participants within S_k appearing after participant i . The equation (2) essentially involves two rounds of Shapley value allocation. The first round is the Shapley value allocation among all possible coalitions, and the second round is the Shapley value allocation within each coalition. It is evident that when the coalition structure is trivial (each forms a separate coalition), the Owen value is equivalent to the Shapley value. Therefore, the Owen value can be considered an extension of the Shapley value under the constraint of a priority coalition structure.

3 Shapley value in ML

The Shapley value, derived initially from cooperative game theory, has gained significant popularity in XAI research. Before introducing the Shapley value in ML, we establish an intuitive understanding of feature attribution through a linear prediction model [17]:

$$\hat{f}(x) = f(x_1, \dots, x_n) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n, \quad (3)$$

where x represents the instance for which its contribution is to be computed. β_j denotes the weight associated with feature x_j . The contribution of the j -th feature to the prediction $\hat{f}(x)$ is given by ϕ_j :

$$\phi_j(\hat{f}) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j), \quad (4)$$

where $E(\beta_j X_j)$ represents the estimated average effect contribution of feature x_j . By summing up the contributions of all features for a given instance, the result is as follows:

$$\begin{aligned} \sum_{j=1}^n \phi_j(\hat{f}) &= \sum_{j=1}^n (\beta_j x_j - E(\beta_j X_j)) \\ &= \hat{f}(x) - E(\hat{f}(X)). \end{aligned} \quad (5)$$

For the non-linear functions more commonly used in ML models, explicit feature coefficients like those in linear models are not available. Fortunately, the Shapley value offers a solution for computing feature attributions. According to equation (1), the calculation of the Shapley value for feature x_i in model \hat{f} is as follows:

$$\phi_i(N, \hat{f}) = \underbrace{\omega(n, s)}_{\text{weight factor}} \underbrace{(\hat{f}(S \cup \{i\}) - \hat{f}(S))}_{\text{marginal contribution}}, \quad (6)$$

where $\hat{f}(S)$ represents the prediction for the set of feature values in S , obtained by integrating over the features that are not included in the set S :

$$\hat{f}(S) = \int \hat{f}(x_S, X_{\bar{S}}) d\mathbb{P}_{X_{\bar{S}}} - E_X(\hat{f}(X)), \quad (7)$$

where x_S represents the observed features in set S , $X_{\bar{S}}$ represents the set of features in the complement, X is random variable in the background dataset, $\mathbb{P}_{X_{\bar{S}}}$ represents the marginal distribution of $X_{\bar{S}}$ over the background dataset. Equation (7) exclusively utilizes marginal integration, assuming the independence of observed features x_S from unobserved features in random variables $X_{\bar{S}}$. In practice, the underlying dependence between features can be accounted for by the conditional distribution $\mathbb{P}_{X_{\bar{S}}/x_S}$. However, the complex dependencies between features and the high dimensionality of data pose challenges in determining the conditional distribution, as investigated in [21].

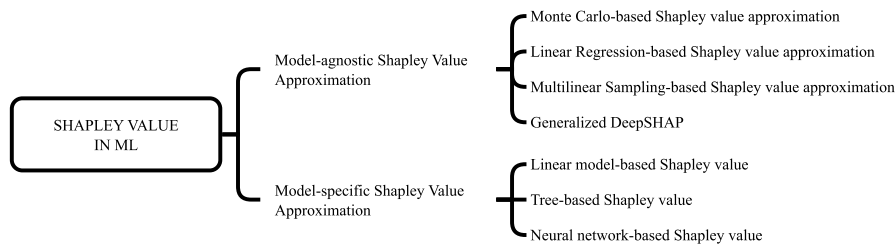


Figure 3 Shapley value estimation methods classification

To illustrate the estimation of Shapley value in ML, a motivation example is shown in Fig. 2(b), which depicts a process for computing feature attribution using the Shapley value. It introduces a background dataset denoted by a single data x^* to replace the missing feature in the explained instance x . It is important to note that, in the context of ML, the value function corresponding to the empty set represents the output expectation of the background dataset, i.e., $v(\emptyset) = \hat{f}(x^*)$ (In this case, a single base value x^* is used to represent the background data). The sum of $v(\emptyset)$ and the feature attributions for the two features of instance x is equal to the predicted value $\hat{f}(x)$ of instance x . It is evident that in linear predictive models \hat{f} , the feature attributions assigned by the Shapley value are consistent with the results obtained from equation (4). The Shapley values attribution for individual instances, as well as the aggregation of Shapley values across multiple instances, offer explanatory insights into the behavior of the ML model at both local and global levels. This visualization, as depicted in Fig. 2(c), provides interpretability for the model [14, 22].

On the other hand, the computational burden is also a significant challenge in calculating the Shapley value. This involves iterating through all subsets S , resulting in a computational complexity of $\mathcal{O}(2^n)$, where n represents the number of feature. Assuming 32 features in the dataset, this would require approximately 17.1 billion enumerations, which is unfeasible. In recent years, numerous research efforts have focused on developing approximation methods to accelerate computations of Shapley value. In the following sections, we will introduce and compare these approximation methods as shown in Fig. 3.

3.1 Model-agnostic Shapley value approximation

Model-agnostic Shapley value approximation refers to approximation methods for the Shapley value that are not dependent on a specific type of ML model. The following section will introduce several popular categories of methods.

3.1.1 Monte Carlo-based Shapley value approximation

Monte Carlo sampling [23] is a probabilistic and statistical simulation method commonly used for approximating solutions. Fatima *et al.* [24] proposed a randomized-based approximation algorithm for computing Shapley values,

thereby demonstrating improved performance in terms of approximation accuracy. Castro *et al.* [25] introduced a polynomial-time method based on sampling theory to estimate the Shapley values. It enables the computation of the value of any coalition in polynomial time. Štrumbelj and Kononenko [26] highlighted the utilization of Monte Carlo integration to expedite convergence, wherein quasi-random sampling techniques can be employed as an alternative to pseudo-random sampling. Mitchell *et al.* [27] utilized sampling from a uniform distribution of permutations to compute estimates of the Shapley value. This type of method typically involves sampling from a dataset using a marginal distribution approach.

3.1.2 Linear regression-based Shapley value approximation

Linear regression is a widely used statistical modeling technique that aims to establish a linear relationship between input and output. Lundberg and Lee [14] proposed KernelSHAP, which combines the idea of LIME with weighted least squares optimization to approximate Shapley values.

Unlike LIME, the kernel used here is non-heuristic. Covert and Lee [28] extended the research on this method. Compared to the original KernelSHAP, they employed a new unbiased and variance reduction estimation method to accelerate the convergence speed of KernelSHAP further. Another method, SGD-Shapley [29], follows a similar approach by sampling subsets. However, it employs the Projection Stochastic Gradient Descent (SGD) method to estimate the Weighted Least Squares solution iteratively. Jethani *et al.* [30] proposed FastSHAP for efficiently estimating Shapley values by computing them within a single forward pass. It employs a stochastic gradient descent approach based on a weighted least squares objective function. This method trains the FastSHAP model by minimizing the weighted squared error between the predicted Shapley values and the ground truth Shapley values.

3.1.3 Multilinear sampling-based Shapley value approximation

To reduce the variance, Okhrati and Lipani [31] proposed a multilinear extension technique. Compared to [25], the

improvement of this method primarily lies in the reduction of variance in estimating the statistical quantity, thereby enhancing the accuracy and time efficiency of estimating Shapley values.

3.1.4 Generalized DeepSHAP

Generalized DeepSHAP [32, 33] is an extension of DeepSHAP. Chen *et al.* [33] introduced a generalized scaling rule to explain a wide range of complex models by propagating attributions. This framework extends DeepSHAP to explain any combination of linear, deep, and tree models. It demonstrates that these population attributions provide improved explanations for models with many features and outperform existing model-agnostic attribution techniques by providing an order-of-magnitude improvement in speed.

In general, model-agnostic Shapley value approximation methods offer greater flexibility, while they come with a significant computational burden.

3.2 Model-specific Shapley value approximation

Model-specific Shapley value approximation methods are designed to efficiently compute Shapley values for different ML models leveraging their specific properties and structures. This section will provide a brief overview of several commonly used model-specific Shapley value approximation methods.

3.2.1 Linear model-based Shapley value

Assuming input features independence, the Shapley values of the linear model can be approximated directly from equation (4). This has been previously proven in [14, 26]. Exact computation of the baseline and marginal Shapley values is feasible with a time complexity that scales linearly with the number of features.

3.2.2 Tree-based Shapley value

TreeSHAP is a Shapley value approximation method designed for tree-based models such as decision trees, random forests, and gradient-boosting trees. These models have a hierarchical structure, and TreeSHAP utilizes this structure for efficient computation. It calculates the Shapley value for each feature in a bottom-up manner and leverages feature-splitting conditions to reduce the computational complexity to an acceptable level. There are two model-specific methods for tree models: Path-based TreeSHAP [34] and Interventional TreeSHAP [35, 36]. Interventional TreeSHAP allows for exact computation of baseline and marginal Shapley values using marginal sampling, with time complexity linear in the tree size and the number of baseline values. Path-based TreeSHAP provides deterministic estimates of conditional Shapley values, but it introduces bias by assuming that the tree model itself can approximate conditional expectations. To estimate Shapley

values for tree ensemble models, both methods compute explanations for each tree and then linearly combine them. This enables exact computation of baseline and marginal Shapley values due to the *Additivity/Linearity* property.

3.2.3 Neural network-based Shapley value

Neural network-based Shapley value is tailored for deep learning models. It leverages the hierarchical structure of neural networks to approximate Shapley values with lower computational complexity. Shrikumar *et al.* [12] proposed DeepLIFT, which approximates Shapley values through recursive multipliers. Given a single reference feature vector x^* and the model's output $y = \hat{f}(x)$, where $\Delta y = \hat{f}(x) - \hat{f}(x^*)$ and $\Delta x_i = x_i - x_i^*$. We have $\Delta y = \sum_{i=1}^n p_{x_i \hat{f}} \Delta x_i$, where $p_{x_i \hat{f}}$ is the multiplier and the approximate Shapley value of x_i is $p_{x_i \hat{f}} \Delta x_i$. Considering a feedforward neural network $v^{NN}(\cdot)$, it represents complex models as compositions of simple functions.

$$v(x) \approx v^{NN}(x) = g_L \circ \dots \circ g_l \circ \dots \circ g_1(x), \quad (8)$$

where $g_l(\cdot) = [g_l^1(\cdot), \dots, g_l^{n_l}(\cdot)]^T$ represents the l -th hidden layer with n neurons, where $g_l^j \in \mathbb{R}^n$ with $j = 1, \dots, n_l$. By applying the chain rule and linear approximation, an approximation of the Shapley value is obtained.

$$\begin{aligned} \varphi_i(N, \hat{f}) &= p_{x_i \hat{f}}(x_i - x_i^*), \\ &= \sum_{j_1=1}^{n_1} p_{x_i g_1^{j_1}} \cdots \sum_{j_{l-1}=1}^{n_{l-1}} p_{g_{l-1}^{j_{l-1}} g_l^{j_l}} \\ &\quad \cdots \sum_{j_L=1}^{n_L} p_{g_{L-1}^{j_{L-1}} g_L^{j_L}} p_{g_{L-1}^{j_{L-1}} g_L^{j_L}} \\ &= \frac{\varphi_i(g_L^j, g_{L-1})}{g_{L-1}^{j_{L-1}} - g_{L-1}^{*j_{L-1}}}. \end{aligned} \quad (9)$$

DeepLIFT explicitly specifies that the base point x^* is a single data point, rather than a data distribution. The choice of x^* depends on the specific application context. For example, in image recognition, it is common to set the values of all three RGB channels to zero as the base point. It is worth noting that the connection between DeepLIFT and Shapley value is established in DeepSHAP [14]. DeepSHAP builds upon DeepLIFT and differs by using marginal distributed data as the base point. In essence, the two methods are equivalent when not considering the choice of base value. Ancona *et al.* [37] introduced a polynomial-time approximation for estimating Shapley values in deep neural networks by incorporating the concept of uncertainty propagation. Wang *et al.* [38] proposed Shapley Explanation Networks (SHAPNETs), which leverages Shapley values as a latent representation for deep

models, enabling hierarchical explanations and explanation regularization. SHAPNETs efficiently compute explanations during training and testing by providing hierarchical explanations during the forward propagation process.

In model-agnostic approximate methods, Kernel SHAP [14] is currently widely used. For tree and neural network models, TreeSHAP [35, 36] and DeepSHAP [14] are more applicable. All of these methods are integrated in the SHAP library, available at: <https://github.com/shap/shap>. Overall, compared to model-agnostic methods, model-specific approximations of Shapley value have greatly improved computational efficiency but lack flexibility.

4 Restricted Shapley value in ML

In ML, there are two primary types of restricted Shapley values: Owen value and causal Shapley value. The Owen value takes into account priority coalition structures, while the causal Shapley value considers causal relationships between features (i.e., causal graphs are not flat) [39]. This section will provide a brief introduction to these two types of restricted Shapley values.

4.1 Owen value in ML

Based on equation (2), it is straightforward to derive the expression for Owen value in ML:

$$Ow_i(N, \hat{f}, C) = \sum_{R \subseteq M \setminus \{i\}} \sum_{T \subseteq S_k \setminus \{i\}} \frac{(|M| - |R| - 1)! |R|! (|S_k| - |T| - 1)! |T|!}{|M|!} \cdot [\hat{f}(R \cup T \cup \{i\}) - \hat{f}(R \cup T)]. \quad (10)$$

Currently, Owen value has not been widely applied in feature attribution research for ML models. However, Owen

value holds great potential and is expected to be employed in various domains of ML models. For instance, it can be leveraged to explain ML-based automatic driving decision models. This is particularly relevant due to the collective nature of traffic vehicles, which naturally forms a priority coalition. In such cases, all input features associated with the vehicle are considered part of this priority coalition. Consequently, Owen value emerges as a more effective attribution method in this context. The approximation methods based on Monte Carlo sampling are also employed in estimating the Owen value [40].

4.2 Causal Shapley value in ML

The Shapley value often overlooks the causal structure within the data, inadvertently diminishing the attribution of the dependent variable in the data. Therefore, this section provides a review of the research on causal Shapley value. To incorporate causality, Frye *et al.* [41] introduced the framework of Asymmetric Shapley values (ASVs), which is an extension of the traditional Shapley values that relax *symmetry* axiom, which can be applied to various ML models without requiring a complete causal graph, even with limited knowledge of causal relationships. Heskes *et al.* [42] proposed causal Shapley value framework, which utilizes Pearl's do-calculus [43] to overcome the assumption of independence and demonstrates how to derive these causal Shapley values for general causal graphs without sacrificing any desirable properties. This approach enables a more accurate and meaningful attribution of the overall effects of features on predictions. To establish a unified framework capturing direct and indirect effects among variables with causal relationships, Wang *et al.* [44] proposed Shapley Flow, which takes

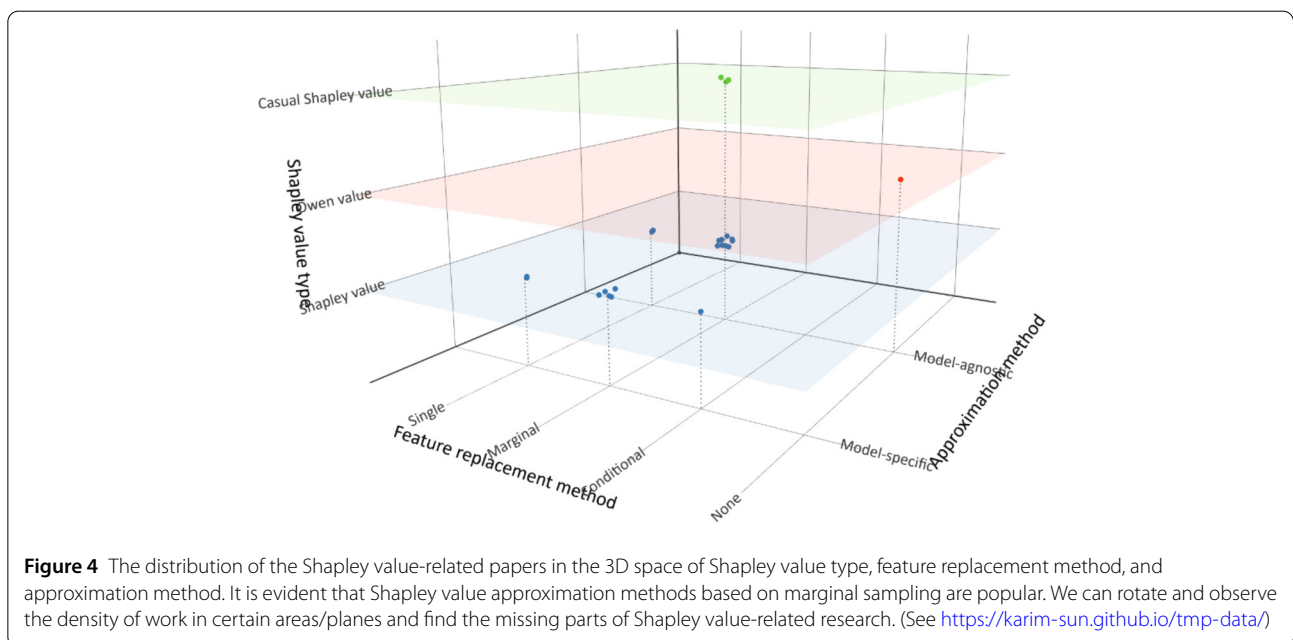


Figure 4 The distribution of the Shapley value-related papers in the 3D space of Shapley value type, feature replacement method, and approximation method. It is evident that Shapley value approximation methods based on marginal sampling are popular. We can rotate and observe the density of work in certain areas/planes and find the missing parts of Shapley value-related research. (See <https://karim-sun.github.io/tmp-data/>)

Table 1 Methods for estimating the Shapley value (SV) and the restricted SV

Reference	SV type	Approximation method	Feature replacement method
Fatima et al. [24]	Shapley value	Monte Carlo	marginal
Castro et al. [25]			marginal
Štrumbelj et al. [26]			marginal
Mitchell et al. [27]			marginal
Lundberg et al. [14]	Shapley value	Linear Regression	marginal
Covert et al. [28]			marginal
Simon et al. [29]			single
Jethani et al. [30]			marginal
Okhrati et al. [31]	Shapley value	Multilinear Sampling	single
Chen et al. [32], Chen et al. [33]	Shapley value	Generalized DeepSHAP	marginal
Lundberg et al. [14], Štrumbelj et al. [26]	Shapley value	Linear model	marginal
Mase et al. [34]	Shapley value	Tree model	conditional
Lundberg et al. [35], Lundberg et al. [36]	Shapley value	Deep model	marginal
Shrikumar et al. [12]			single
Lundberg et al. [14]			marginal
Wang et al. [38]			single
Saavedra-Nieves et al. [40]	Owen value	Monte Carlo	None
Frye et al. [41]	Causal Shapley value	Monte Carlo	marginal
Heskes et al. [42]			marginal
Wang et al. [44]			marginal

We ranked the algorithms based on Shapley value, Owen value, and causal Shapley value. Approximation methods for Shapley value were categorized into model-specific and model-agnostic approaches. Model-agnostic methods were further divided into Monte Carlo sampling, linear regression, multilinear sampling, and generalized DeepSHAP. Model-specific approaches were categorized based on linear models, tree models, and deep neural network models. Both Owen value and causal Shapley value, introduced in our study, were model-agnostic and approximated using Monte Carlo sampling. Additionally, the three types of feature replacement methods were represented as marginal, conditional, and single, which were used for substituting missing features. ‘None’ indicates that is not relevant within the scope of ML.

into account the entire causal graph and allocates credit to edges instead of treating nodes as the fundamental unit of credit allocation. By constructing the flow of Shapley values in directed acyclic graphs, it can visualize both the direct and indirect influences of variables. With the development of causal theory, modeling causal relationships has made significant progress [43]. However, when it comes to real-world problems, the precise quantification of causal relationships remains a major challenge that hinders the practical application of causal Shapley.

Based on the contents described in Sects. 3 and 4, we classify the methods into a three-dimensional framework based on the dimensions of Shapley value type, feature replacement method, and approximation method. This can be visualized in a 3D view, showcasing the distribution of existing Shapley value-based feature attribution papers (Fig. 4 provides only a 2D snapshot, and we encourage readers to access the interactive online version for a better visualization). Table 1 serves as another representation of all the Shapley value methods, facilitating quick navigation.

5 Application

The application of Shapley value permeates three stages of ML modeling: pre-modeling, mid-modeling, and post-modeling. In the pre-modeling stage, it is utilized for fea-

ture selection. In the mid-modeling stage, it is employed for credit assignment in cooperative multi-agent reinforcement learning (MARL). In the post-modeling stage, it is used for data valuation and explaining model. The following comprehensive review will discuss these applications in detail.

5.1 Feature selection

Feature selection is a process of identifying relevant features from a dataset, aiming to improve model performance and reduce complexity. It can be seen in Fig. 2(c), where *c2* displays the feature importance ranking calculated based on Shapley values in a cervical cancer risk prediction model. This assists users in selecting important features for modeling. Cohen et al. [45] presented a feature selection algorithm based on Shapley values, called Contribution-Selection Algorithm (CSA). The study utilizes uniform sampling of feature subsets to approximate the Shapley values, which are then used to quantify the contribution of each feature to the classification task. By iteratively evaluating the usefulness of features based on their Shapley value contributions, the algorithm is able to select the most relevant features accordingly. In [46–48], feature selection is also explored with the core idea of ranking and selecting features based on the absolute magnitude of their Shapley values.

5.2 Credit assignment in cooperative MARL

Cooperative MARL is a reinforcement learning approach that investigates how multiple agents can collaborate to solve problems, requiring them to take joint actions to achieve shared objectives. However, MARL faces the credit assignment problem, which pertains to how to fairly distribute the global reward among individual agents. In this context, Wang *et al.* [49] proposed a local reward method called Shapley Q Value Deep Deterministic Policy Gradient (SQDDPG). The SQDDPG algorithm is evaluated in cooperative navigation, predator-prey, and traffic intersection scenarios. Experimental results demonstrate significant improvements in convergence speed. Furthermore, Wang *et al.* [50] proposed Shapley Q learning (SHAQ), which extends the application of Shapley value theory to Markov convex games, referred to as Markov Shapley value (MSV), and employs it as a value decomposition method for global reward games. While the aforementioned two works claim to have achieved promising results, their performance does not exhibit significant superiority over other cooperative MARL methods. This limitation may stem from the insufficient representation capability of the value decomposition model based on Shapley values in capturing the contributions of multiple agents.

5.3 Data valuation

Data valuation is the process of assessing the importance of each sample or instance for predicting the model's outcomes. It focuses on the contribution of individual data points, representing their influence or information value to the model. Ghorbani and Zou [51] proposed a fair data valuation method based on the Shapley measure, which is used to estimate the data Shapley values. These values quantify the individual data's worth in predictions of ML models. Pandl *et al.* [52] investigated how to ensure trustworthy ML in healthcare. By analyzing the suitability of different data valuation methods in the context of medical image classification tasks, particularly in the detection of pleural effusion, it is discovered that the Shapley valuation scheme based on the k -nearest neighbors classifier can successfully value a large number of data instances. Tang *et al.* [53] pointed out that data with low Shapley values often contained mislabeled samples, while data with high Shapley values were more likely to include valuable data points for pneumonia detection. These findings suggest that by analyzing Shapley values, it is possible to identify low-quality samples and samples that are valuable for pneumonia detection within the dataset. Additionally, the potential applications of data valuation methods are demonstrated in terms of incentivizing data sharing, detecting mislabeled data, and protecting private information.

5.4 Explaining model

Explaining model refers to the process of providing an understanding of the underlying basis and logic behind

the model's decision when explaining its predictions. Figure 2(c) displays the Shapley Value Summary Plot, where fewer years of using hormonal contraceptives will decrease the predicted cancer risk, while more years will increase the risk. This can elucidate the model's decision-making mechanism, aiding users in assessing the model's validity. Heuillet *et al.* [54] utilized the Shapley value to quantify the contributions of each agent in a cooperative MARL environment. This approach provides a quantitative explanation, enabling a better understanding of the behavior and decision-making process of agents. Lundberg *et al.* [22] presented Prescience, an ML approach based on an ensemble model, for predicting the near-term risk of hypoxemia during anesthesia care and explaining the specific factors related to patients and surgeries that contribute to this risk. To provide real-time explanations, it has developed an effective and theoretically sound ML technique for explaining the importance of individual features in the model's predictions for each patient. These explanations are presented in concise visualizations for anesthesiologists' use. This capacity to offer simplified explanations removes the typical trade-off between accuracy and interpretability, thereby widening the applicability of ML in the medical field. Similar explanations have already been extensively applied in domains such as finance, autonomous driving, image recognition, and other fields. These applications have successfully utilized explanations based on Shapley value to gain valuable insights and enhance understanding in these domains.

6 Limitation and future research directions

Despite the widespread application of the Shapley value in ML, it still has several limitations. This section provides a comprehensive overview of the limitations of the Shapley value in ML and offers insights into future research directions.

6.1 Limitations

Some of the limitations and challenges associated with the application of Shapley value in ML include: 1. Computational Complexity: Although efforts have been made to develop fast approximation methods to alleviate the computational costs of computing Shapley values, it remains a computationally expensive task, especially when dealing with large-scale models and datasets. 2. Ambiguity in Feature Interactions: Shapley value assumes that contributions of individual features are independent of their interactions with other features. However, in reality, features may interact with each other in complex ways, leading to challenges in accurately attributing contributions. 3. Model Sensitivity: Shapley value can be sensitive to changes in the model or training data. Slight variations in the model or dataset may result in significant changes in the computed Shapley values, making them less stable and

reliable. 4. Interpretability: While the Shapley value provides insights into feature importance, it may not always provide intuitive explanations for complex machine learning models. Understanding the meaning and implications of Shapley values in complex models can be challenging for non-experts [55]. 5. Sample Representativeness: Shapley value relies on sampling techniques to estimate feature contributions. The accuracy of the estimates depends on the representativeness of the sampled data, and biased or unrepresentative samples may lead to inaccurate or misleading results. 6. Axiom Violation: Some approximation methods for Shapley value are based on certain assumptions, which may sacrifice axioms of Shapley value.

6.2 Future research direction

Despite the aforementioned limitations, Shapley value remains the most theoretically comprehensive method for feature attribution in current research. As a result, the future research prospects for Shapley value are still promising.

6.2.1 Model diagnosis

Model diagnosis is primarily used for evaluating and analyzing ML models to identify the model's weaknesses and potential areas for improvement. It involves a comprehensive examination of the model's performance and behavior. Ghorbani *et al.* [56] utilized the Shapley value to quantify the contribution of individual neurons to the predictions and performance of deep neural networks. This provides insights into filters responsible for biased predictions and susceptibility to adversarial attacks in image recognition tasks. By removing specific responsible neurons, the model can achieve fairer predictions for specific subgroups or increased robustness against adversarial attacks. Li *et al.* [57] analyzed erroneous predictions of an automated lane change prediction model using the Shapley value. It reveals that different types of erroneous predictions can be attributed to 1) the model's failure to learn the correct representation of the input space and 2) the interaction between features, resulting in unexpected model behavior. Applying the Shapley value in model diagnosis offers insights into model anomalies for operators or designers, making it a promising auxiliary tool for effective model design. Taking the next step beyond using Shapley value solely for feature attribution is crucial for further advancing model diagnosis in ML.

6.2.2 Model optimization

Shapley value-based feature attribution methods hold the potential to improve model performance by incorporating prior knowledge into the model training process. This can be achieved by assigning higher weights or importance to certain features based on their relevance to the problem domain. By considering domain-specific knowledge, such as expert opinions, domain rules, or known

causal relationships, the attribution results can better align with human intuition. Another approach is to develop hybrid methods that combine Shapley value-based attribution with other training techniques. For instance, incorporating prior knowledge as regularization terms or constraints in the model optimization process can guide the learning process and promote the discovery of more meaningful and interpretable feature attributions. Indeed, there have been previous studies exploring this direction of incorporating prior knowledge into other feature attribution methods. Erion *et al.* [58] demonstrated that encoding human attribution priors as feature attributions can help deep learning models achieve better performance on image classification, gene expression, and healthcare datasets. By incorporating prior knowledge in the form of feature attributions, the models can align with human understanding and exhibit improved performance in various domains. Rieger *et al.* [59] introduced Contextual Decomposition Explanation Penalization, which allows the integration of domain knowledge into the model to mitigate spurious correlations, i.e., shortcut learning, and correct errors. Therefore, encoding prior knowledge through Shapley value during the model training to guide model optimization is a promising approach with significant potential.

7 Conclusion

This review explores the concept of Shapley values and their significance in improving interpretability in ML. It delves into the foundational theory of Shapley values, which originated from cooperative game theory, and discusses their desirable properties. A framework is presented to enhance comprehension and facilitate the identification of relevant algorithms, classifying existing feature attribution methods based on Shapley values across three dimensions: Shapley value type, feature replacement method, and approximation method. A visualization link is provided to illustrate the distribution of these methods in three-dimensional space. Furthermore, the practical application of Shapley values throughout various stages of ML model development is emphasized, including pre-modeling, in-modeling, and post-modeling phases.

From the perspective of Shapley value approximation methods, this study examines several limitations associated with them. These limitations encompass the computational complexity of computing Shapley values, the ambiguity of feature interactions, the sensitivity of models to changes, the challenges in interpretability, the representativeness of sampled data, and the potential violation of axioms in approximation methods. It is crucial to consider these factors. However, despite these limitations, the Shapley value holds significant potential in model diagnosis and performance improvement. Further exploration is

warranted to address various challenges in machine learning, including model defects and shortcut learning. Moreover, selecting appropriate base values to construct explanations that better align with human intuition is also an intriguing avenue to explore.

Acknowledgements

We express our sincere appreciation to Zhihao Cui for his diligent efforts in sourcing relevant research materials. Furthermore, our thanks extend to Yulei Wang for skillfully enhancing the article's structure.

Funding

This work was supported by the National Key Research and Development Program of China under Grant No. 2020AAA0108101, and the National Natural Science Foundation of China under Grant No. U1964201.

Data availability

Not applicable.

Declarations

Competing interests

Prof. Hong Chen and Yanjun Huang are editorial board members for *Autonomous Intelligent Systems* and were not involved in the editorial review or the decision to publish this article. All authors declare that there are no other competing interests.

Author contributions

ML and HS contributed to problem formulation, discussion of ideas, and the collection and organization of the paper. HC and YH contributed to structuring the article. All authors have read and approved the final manuscript.

Author details

¹College of Electronics and Information Engineering, Tongji University, Shanghai, 201804, China. ²College of Automotive Studies, Tongji University, Shanghai, 201804, China. ³Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai, 201203, China.

Received: 7 November 2023 Revised: 21 December 2023

Accepted: 26 December 2023 Published online: 09 February 2024

References

1. D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot et al., Mastering the game of go with deep neural networks and tree search. *Nature* **529**(7587), 484–489 (2016)
2. D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton et al., Mastering the game of go without human knowledge. *Nature* **550**(7676), 354–359 (2017)
3. O. Vinyals, I. Babuschkin, W.M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D.H. Choi, R. Powell, T. Ewalds, P. Georgiev et al., Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature* **575**(7782), 350–354 (2019)
4. P.R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T.J. Walsh, R. Capobianco, A. Devlic, F. Eckert, F. Fuchs et al., Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature* **602**(7896), 223–228 (2022)
5. S. Feng, H. Sun, X. Yan, H. Zhu, Z. Zou, S. Shen, H.X. Liu, Dense reinforcement learning for safety validation of autonomous vehicles. *Nature* **615**(7953), 620–627 (2023)
6. A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Židek, A.W. Nelson, A. Bridgland et al., Improved protein structure prediction using potentials from deep learning. *Nature* **577**(7792), 706–710 (2020)
7. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017). <https://doi.org/10.48550/arXiv.1706.03762>. [arXiv:1706.03762](https://arxiv.org/abs/1706.03762)
8. I. Zliobaite, A survey on measuring indirect discrimination in machine learning. *arXiv preprint* (2015). [arXiv:1511.00148](https://arxiv.org/abs/1511.00148)
9. P. Regulation, General data protection regulation. *Intouch* **25**, 1–5 (2018)
10. M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in *International Conference on Machine Learning* (PMLR, 2017), pp. 3319–3328
11. S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**(7), 0130140 (2015)
12. A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in *International Conference on Machine Learning* (PMLR, 2017), pp. 3145–3153
13. M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 1135–1144
14. S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30** (2017). <https://doi.org/10.48550/arXiv.1705.07874>. [arXiv:1705.07874](https://arxiv.org/abs/1705.07874)
15. B. Rozemberczki, L. Watson, P. Bayer, H.-T. Yang, O. Kiss, S. Nilsson, R. Sarkar, The Shapley value in machine learning. *arXiv preprint* (2022). [arXiv:2202.05594](https://arxiv.org/abs/2202.05594)
16. H. Chen, I.C. Covert, S.M. Lundberg, S.-I. Lee, Algorithms to estimate Shapley value feature attributions. *Nat. Mach. Intell.* **5**, 590–601 (2023)
17. C. Molnar, *Interpretable Machine Learning*. [Lulu.com](https://lulua.com) (2020)
18. R. Branzei, D. Dimitrov, S. Tijs, *Models in Cooperative Game Theory*, vol. 556 (Springer, Berlin, 2008)
19. H.W. Kuhn, A.W. Tucker, *Contributions to the Theory of Games*, vol. 28 (Princeton University Press, Princeton, 1953)
20. G. Owen, Values of games with a priori unions, in *Mathematical Economics and Game Theory: Essays in Honor of Oskar Morgenstern* (Springer, Berlin, 1977), pp. 76–88
21. C. Frye, D. Mijolla, T. Begley, L. Cowton, M. Stanley, I. Feige, Shapley explainability on the data manifold. *arXiv preprint* (2020). [arXiv:2006.01272](https://arxiv.org/abs/2006.01272)
22. S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D.E. Liston, D.K.-W. Low, S.-F. Newman, J. Kim et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**(10), 749–760 (2018)
23. F. James, Monte Carlo theory and practice. *Rep. Prog. Phys.* **43**(9), 1145 (1980)
24. S.S. Fatima, M. Wooldridge, N.R. Jennings, A linear approximation method for the Shapley value. *Artif. Intell.* **172**(14), 1673–1699 (2008)
25. J. Castro, D. Gómez, J. Tejada, Polynomial calculation of the Shapley value based on sampling. *Comput. Oper. Res.* **36**(5), 1726–1730 (2009)
26. E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **41**, 647–665 (2014)
27. R. Mitchell, J. Cooper, E. Frank, G. Holmes, Sampling permutations for Shapley value estimation. *J. Mach. Learn. Res.* **23**, 1–46 (2022)
28. I. Covert, S.-I. Lee, Improving kernelshap: practical Shapley value estimation using linear regression, in *International Conference on Artificial Intelligence and Statistics* (PMLR, 2021), pp. 3457–3465
29. G. Simon, T. Vincent, A projected stochastic gradient algorithm for estimating Shapley value applied in attribute importance, in *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4* (Springer, Berlin, 2020), pp. 97–115
30. N. Jethani, M. Sudarshan, I.C. Covert, S.-I. Lee, R. Ranganath, Fastshap: real-time Shapley value estimation, in *International Conference on Learning Representations* (2021)
31. R. Okhrati, A. Lipani, A multilinear sampling algorithm to estimate Shapley values, in *2020 25th International Conference on Pattern Recognition (ICPR)* (IEEE, New York, 2021), pp. 7992–7999
32. H. Chen, S. Lundberg, S.-I. Lee, Explaining models by propagating Shapley values of local components, in *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability* (2021), pp. 261–270
33. H. Chen, S.M. Lundberg, S.-I. Lee, Explaining a series of models by propagating Shapley values. *Nat. Commun.* **13**(1), 4512 (2022)
34. M. Mase, A.B. Owen, B. Seiler, Explaining black box decisions by shapley cohort refinement. *arXiv preprint* (2019). [arXiv:1911.00467](https://arxiv.org/abs/1911.00467)
35. S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**(1), 56–67 (2020)
36. S.M. Lundberg, G.G. Erion, S.-I. Lee, Consistent individualized feature attribution for tree ensembles. *arXiv preprint* (2018). [arXiv:1802.03888](https://arxiv.org/abs/1802.03888)

37. M. Ancona, C. Oztireli, M. Gross, Explaining deep neural networks with a polynomial time algorithm for Shapley value approximation, in *International Conference on Machine Learning* (PMLR, 2019), pp. 272–281
38. R. Wang, X. Wang, D.I. Inouye, Shapley explanation networks. arXiv preprint (2021). [arXiv:2104.02297](https://arxiv.org/abs/2104.02297)
39. D. Janzing, L. Minorics, P. Blöbaum, Feature relevance quantification in explainable AI: a causal problem, in *International Conference on Artificial Intelligence and Statistics* (PMLR, 2020), pp. 2907–2916
40. A. Saavedra-Nieves, I. García-Jurado, M.G. Fiestras-Janeiro, Estimation of the Owen value based on sampling, in *The Mathematics of the Uncertain: A Tribute to Pedro Gil* (2018), pp. 347–356
41. C. Frye, C. Rowat, I. Feige, Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *Adv. Neural Inf. Process. Syst.* **33**, 1229–1239 (2020)
42. T. Heskes, E. Sijben, I.G. Bucur, T. Claassen, Causal Shapley values: exploiting causal knowledge to explain individual predictions of complex models. *Adv. Neural Inf. Process. Syst.* **33**, 4778–4789 (2020)
43. J. Pearl, *Causality* (Cambridge University Press, Cambridge, 2009)
44. J. Wang, J. Wiens, S. Lundberg, Shapley flow: a graph-based approach to interpreting model predictions, in *International Conference on Artificial Intelligence and Statistics* (PMLR, 2021), pp. 721–729
45. S. Cohen, E. Ruppín, G. Dror, Feature selection based on the Shapley value. *Other Words* **1**(98Eq), 155 (2005)
46. W.E. Marcílio, D.M. Eler, From explanations to feature selection: assessing shap values as feature selection mechanism, in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (IEEE, New York, 2020), pp. 340–347
47. A. Gramegna, P. Giudici, Shapley feature selection. *FinTech* **1**(1), 72–80 (2022)
48. M. Zaeri-Amirani, F. Afghah, S. Mousavi, A feature selection method based on Shapley value to false alarm reduction in icus a genetic-algorithm approach, in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, New York, 2018), pp. 319–323
49. J. Wang, Y. Zhang, T.-K. Kim, Y. Gu, Shapley q-value: a local reward approach to solve global reward games, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34 (2020), pp. 7285–7292
50. J. Wang, Y. Zhang, Y. Gu, T.-K. Kim, Shaq: incorporating Shapley value theory into multi-agent q-learning. *Adv. Neural Inf. Process. Syst.* **35**, 5941–5954 (2022)
51. A. Ghorbani, J. Zou, Data Shapley: equitable valuation of data for machine learning, in *International Conference on Machine Learning* (PMLR, 2019), pp. 2242–2251
52. K.D. Pandl, F. Feiland, S. Thiebes, A. Sunyaev, Trustworthy machine learning for health care: scalable data valuation with the Shapley value, in *Proceedings of the Conference on Health, Inference, and Learning* (2021), pp. 47–57
53. S. Tang, A. Ghorbani, R. Yamashita, S. Rehman, J.A. Dunnmon, J. Zou, D.L. Rubin, Data valuation for medical imaging using Shapley value and application to a large-scale chest X-ray dataset. *Sci. Rep.* **11**(1), 1–9 (2021)
54. A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Collective explainable AI: explaining cooperative strategies and agent contribution in multiagent reinforcement learning with Shapley values. *IEEE Comput. Intell. Mag.* **17**(1), 59–71 (2022)
55. I.E. Kumar, S. Venkatasubramanian, C. Scheidegger, S. Friedler, Problems with Shapley-value-based explanations as feature importance measures, in *International Conference on Machine Learning* (PMLR, 2020), pp. 5491–5500
56. A. Ghorbani, J.Y. Zou, Neuron Shapley: discovering the responsible neurons. *Adv. Neural Inf. Process. Syst.* **33**, 5922–5932 (2020)
57. M. Li, Y. Wang, H. Sun, Z. Cui, Y. Huang, H. Chen, Explaining a machine-learning lane change model with maximum entropy Shapley values, in *IEEE Transactions on Intelligent Vehicles* (2023)
58. G. Erion, J.D. Janizek, P. Sturmfels, S.M. Lundberg, S.-I. Lee, Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nat. Mach. Intell.* **3**(7), 620–631 (2021)
59. L. Rieger, C. Singh, W. Murdoch, B. Yu, Interpretations are useful: penalizing explanations to align neural networks with prior knowledge, in *International Conference on Machine Learning* (PMLR, 2020), pp. 8116–8126

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)